ISSN 2161-3974
Full-Length Paper

Journal of eScience Librarianship 12(1): e628
DOI: https://doi.org/10.7191/jeslib.628

Journal of eScience Librarianship
putting the pieces together: theory and practice

Show me the data! Data sharing practices demonstrated in published research at the University of Massachusetts Chan Medical School

Tess Grynoch, University of Massachusetts Chan Medical School, Worcester, MA, US, tess.grynoch@umassmed.edu
Kimberly MacKenzie, University of Massachusetts Chan Medical School, Worcester, MA, US

Abstract

Objective: In the interest of making data findable, accessible, interoperable, and reusable (FAIR), the National Institutes of Health (NIH) will institute a new Data Management and Sharing Policy in January 2023. This policy will require researchers applying for NIH funding to submit a Data Management and Sharing Plan. As 63% of grant dollars received by University of Massachusetts Chan Medical School (UMass Chan) researchers comes from the NIH, we explored whether UMass Chan researchers are currently sharing data associated with their published research and how they shared their data.

Methods: PubMed was searched for articles published in 2019 with a UMass Chan researcher as either the first or last author. These articles were examined for evidence of original or reused data, the type of data, whether the article stated that data was available, and where and how to find that data.

Results: Of the 361 articles with original data, 26% had a data availability statement. However, most articles (71%) did not mention where data could be accessed. The data storage location of the estimated 1551 original datasets was similarly not mentioned for 74% the datasets with the next largest category being available upon request (8.6%). Genomic data repositories such as the Gene Expression Omnibus were among the top repositories used by authors. Similar areas for improvement were noted for permanent identifier use (46% had a permanent identifier), using non-proprietary file formats (most popular format was Excel), and citing reused data. Authors who published open access were more likely to share their data.

Received: July 9, 2022 **Accepted**: November 4, 2022 **Published**: February 15, 2023

Competing Interests: The authors declare that they have no competing interests.

Data Availability: The data and code are available in eScholarship@UMassChan: Grynoch, Tess and Kimberly MacKenzie. 2022. "Data and Code from 'Show Me the Data! Data Sharing Practices Demonstrated in Published Research at the University of Massachusetts Chan Medical School.'" [Data set and code]. eScholarship@UMassChan. https://doi.org/10.13028/BPQ6-HF10.

Conclusions: While some researchers at UMass Chan have embraced data sharing, particularly genomic data sharing, we expect there will be more data shared in the coming years with the implementation of the new NIH Data Management and Sharing Policy.

Introduction

"Data sharing" is an amorphous term. While the sharing of research data is now required by some funders and journals, they often lack clear guidance as to what constitutes data sharing, and many roadblocks remain. These challenges include the time and financial cost to the preparation and storage of sharable data, as well as the fear of being "scooped" (Tenopir et al. 2011). While the FAIR principles proposed in 2015 promoted data sharing and reuse, under the tenants of making data Findable, Accessible, Interoperable, and Reusable, it is unclear the extent to which these principles have been embraced by the biomedical research community (Wilkinson et al. 2016). In Figshare's 2019 State of Open Data Survey, 52% of frequent data-sharers had never heard of the FAIR principles (Digital Science et al. 2019). Yet, the survey had a four-fold increase in number of responses as compared to previous years which attests to the growing importance of open data. The growth of open data mirrors the rise of the open science movement with an increasing number of authors publishing their work open access (Piwowar et al. 2018) and making other outputs openly available. Many open access publishers have also adopted data sharing policies. Whether or not publishing open access is correlated with data sharing is still unclear.

The National Institutes of Health (NIH) first instituted a data sharing policy in 2003. Under this policy, the "NIH expects that data be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data" (National Institutes of Health 2022a). This policy was very broad in scope, due to the variety of research and data collected in NIH-funded research, and lacked guidelines on "specific ways of documenting, formatting, presenting, or transporting data" (National Institutes of Health 2022a). The NIH offered more concrete guidelines in 2014, with the NIH Genomic Data Sharing Policy, which requires sharing of "large-scale human or non-human genomic data" (National Institutes of Health 2014). The NIH offers specific guidelines and examples of when and how genomic data should be shared (National Institutes of Health 2022b), as well as hosts and supports multiple repositories for researchers to make use of (National Institutes of Health 2022c). In 2023, the new NIH Data Management and Sharing Policy will go into effect, requiring all grant proposals to include a data management and sharing plan and research data to be shared as much as possible (National Institutes of Health 2020b). Non-compliance with the new policy could affect future funding.

The NIH funded approximately $41.7 billion in medical research in 2020 (National Institutes of Health 2020a). The University of Massachusetts Chan Medical School (UMass Chan) received $290.9 million in NIH funding that same year (UMass Medical School Communications 2021). As the new data management

and sharing mandate will affect most researchers the Lamar Soutter Library at UMass Chan supports, as it will for all medical and research universities, we, as data librarians, need to understand the current data sharing practices of our researchers in order to best plan future support offerings. This study examined whether researchers at UMass Chan Medical School were sharing their data and, if so, where and how those data are being shared.

Methods

Data Acquisition

For the initial data acquisition, we searched PubMed on September 10, 2020, for all articles published by University of Massachusetts Chan Medical School authors in 2019. UMass Chan was known as the University of Massachusetts Medical School at that time and the search string includes known abbreviations and associated centers (Table 1). Search results were downloaded as a .csv file and only those with a publication date of 2019 listed in the publication date column were uploaded into REDCap, electronic data capture tool hosted at UMass Chan (Harris et al. 2009). REDCap then parsed the PubMed metadata into individual records and pre-filled the citation information section of our data collection form.

Table 1: PubMed Search String for the University of Massachusetts Medical School

(Massbiologics[Affiliation] OR "Commonwealth Medicine"[Affiliation] OR "Meyers Primary Care Institute"[Affiliation] OR ("E. K. Shriver Center"[Affiliation] OR "Eunice Kennedy Shriver Center"[Affiliation]) OR ("University of Massachusetts Medical School"[ad] OR "UMass Medical School"[ad] OR "UMass Memorial"[ad] OR "University of Massachusetts Worcester"[ad] OR "University of Massachusetts Medical Center"[ad] OR UMMS[ad] OR "University of Massachusetts School of Medicine"[ad]) NOT review[pt] NOT letter[pt] NOT news[pt] NOT editorial[pt]) AND 2019[dp]

Using the PubMed ID number included with each record, we first examined the PubMed record to determine whether an article would be included in the final sample. To be included, the first and/or last author of the article must have listed their affiliation as UMass Chan. We chose this limit as it is generally the first author who serves as corresponding author, responsible for complying with publisher policies, while the last author is most often the principal investigator funding the research and responsible for complying with funder mandates. Beyond authorship, articles must have produced original data or reused existing data. Commentaries, editorials, and reviews were excluded from analysis. First and/or last authorship and the presence of original or reused data were all noted in the REDCap data collection form.

The articles selected for inclusion were then examined individually and the following information was recorded in the REDCap data collection form:

- whether and where data availability was referenced in the article,
- who the research funders were, and
- whether the article was published open access. Articles openly accessible through PubMed Central only were not considered as published open access.

For articles with original research data, we made note of

- the number of data sets,
- the type of data,
- where the data was stored,
- whether the data had a permanent identifier (e.g., a DOI),
- the data file formats, and
- the license for reuse.

For articles with reused data, we noted whether the data was the author's or whether they used someone else's data in their research. For this analysis we considered clinical data, such as patient health record information, as someone else's data and not belonging to an author. As UMass Chan is a medical school and associated with a research hospital, many articles focused on patient health information.

Statistical Methods

Data cleaning and statistical analysis were performed in R (version 4.0.2). All Chi-square tests performed were Pearson's Chi-square tests with Yates' continuity correction. The data and code are available in eScholarship@UMassChan, doi: 10.13028/bpq6-hf10.

Results

Of the 1,357 articles published in 2019 found in PubMed, 535 research articles were included in the final analysis (see Figure 1 for the how articles were chosen). Of the 535 articles chosen, 314 articles included original data only, 174 included reused data only, and 47 included both. More than 300 articles included a branch of the NIH as a research funder.

Articles with Original Data

Of the 361 articles that contained original research data, 256 (71%) did not mention if or where the data could be accessed (Figure 2). A data availability statement was included in 26% of articles that covered how to access at least a portion of the data, while other articles mentioned data access in supplemental files, methods sections, acknowledgements, notes, references, and other locations.

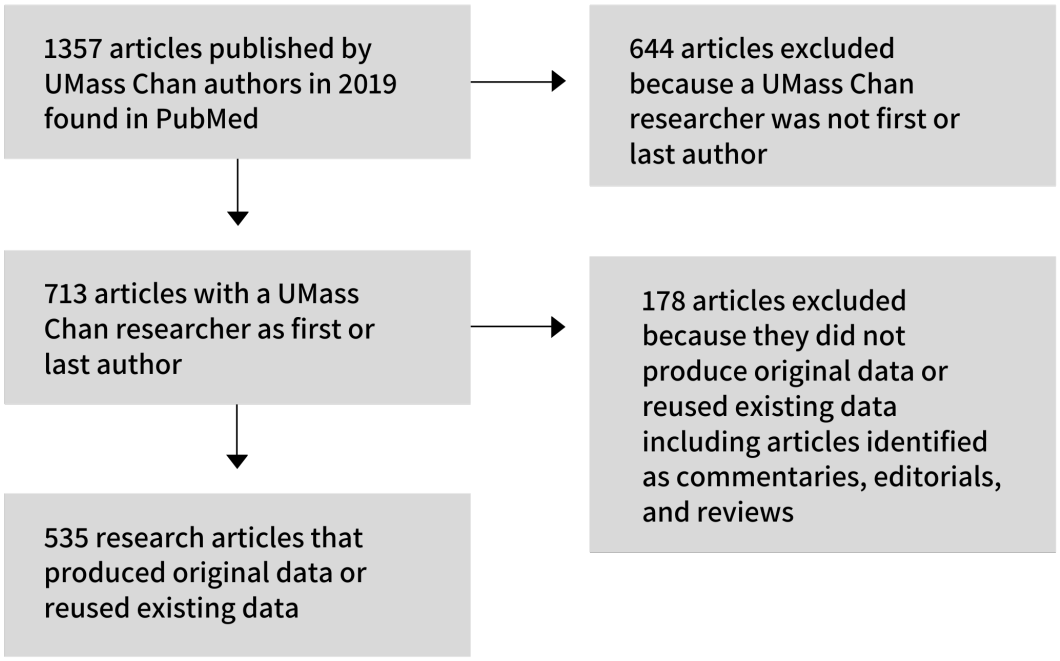

Figure 1: Flow chart of inclusion criteria for articles that were fully reviewed for the study.

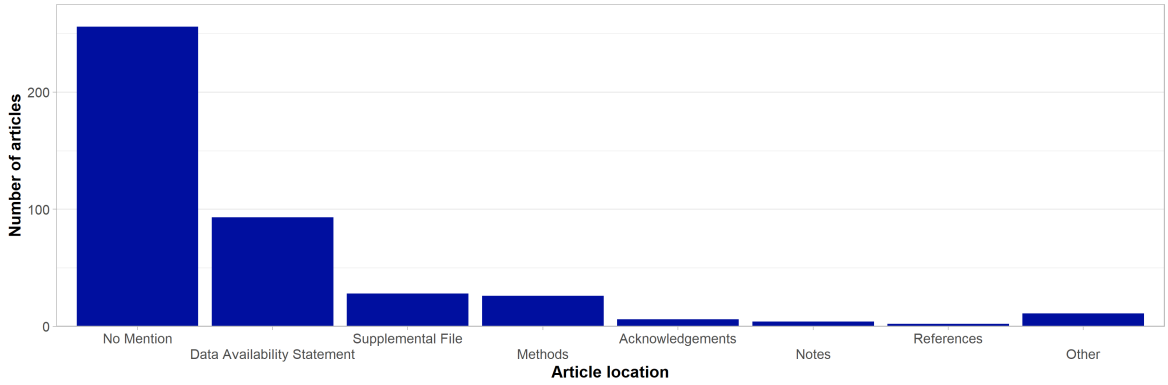

Figure 2: Classification of the 361 articles with original data surveyed by the section within the article where original data availability is mentioned. When data availability was mentioned in multiple locations, it was counted in each location. 71% of the articles did not provide any mention of original data availability.

From the 361 articles with original data, we estimated 1551 potential original datasets. Of these, 1156 (74%) had no data storage location mentioned, 124 (8.0%) were stored in repositories, 106 (6.8%) in a supplemental file on the journal website, and 133 (8.6%) were listed as data being "available upon reasonable request" (Figure 3).

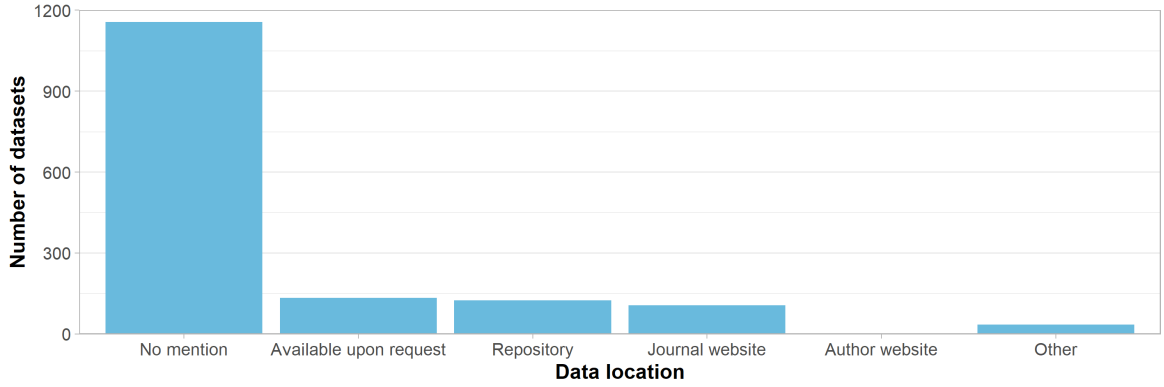

Figure 3: Classification of the 1551 original datasets by the location the dataset was shared as mentioned in the articles. 74% had no data storage location mentioned.

Genomic data was by far the type of data shared most often, and GEO, the Gene Expression Omnibus, was the data repository mentioned most often. Other genomic databases are included in the top 6 repositories mentioned (Figure 4), including the NCBI Sequence Read Archive and the Protein Data Bank.

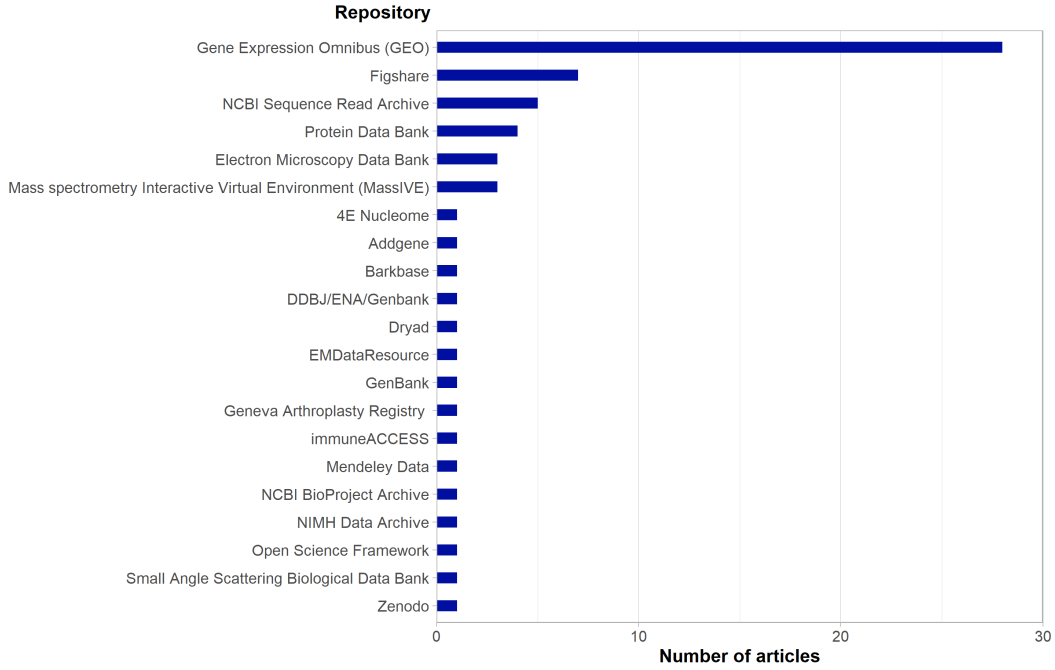

Figure 4: Data repositories used by UMass Chan researchers and the number of articles that had data deposited in that repository.

Of the 266 datasets stored in repositories, on the author's website, in a supplemental file, or listed as being stored elsewhere, 54% did not have an associated identifier (e.g., an ID number, a DOI). The most commonly used license was CC-BY, by 37%. Other Creative Commons licenses were used, but for 31.6%, the type of license was unclear.

Data formats

The vast majority of data files shared were Excel spreadsheets (86 files), followed by text format and PDF. A much smaller number shared as .csv formatted spreadsheets, a non-proprietary format.

Articles with Reused Data

Of the 221 articles with reused data, only 14% were from an author reusing their own data. Like the articles with original data, the majority of articles with reused data (71%) do not mention data availability (Figure 5). When data availability was mentioned, it was most often in the Methods section (12%) or data availability statement (10%).

Figure 5: Classification of the 221 articles with reused data surveyed by the section within the article where reused data availability is mentioned. When data availability was mentioned in multiple locations, it was counted in each location. 71% of articles do not mention where to access the reused data.

Open Access

182 articles were published open access and open access publishing was correlated with a mention of data sharing within the article (χ 2 = 19.626, df = 1, p<0.001) (Figure 6).

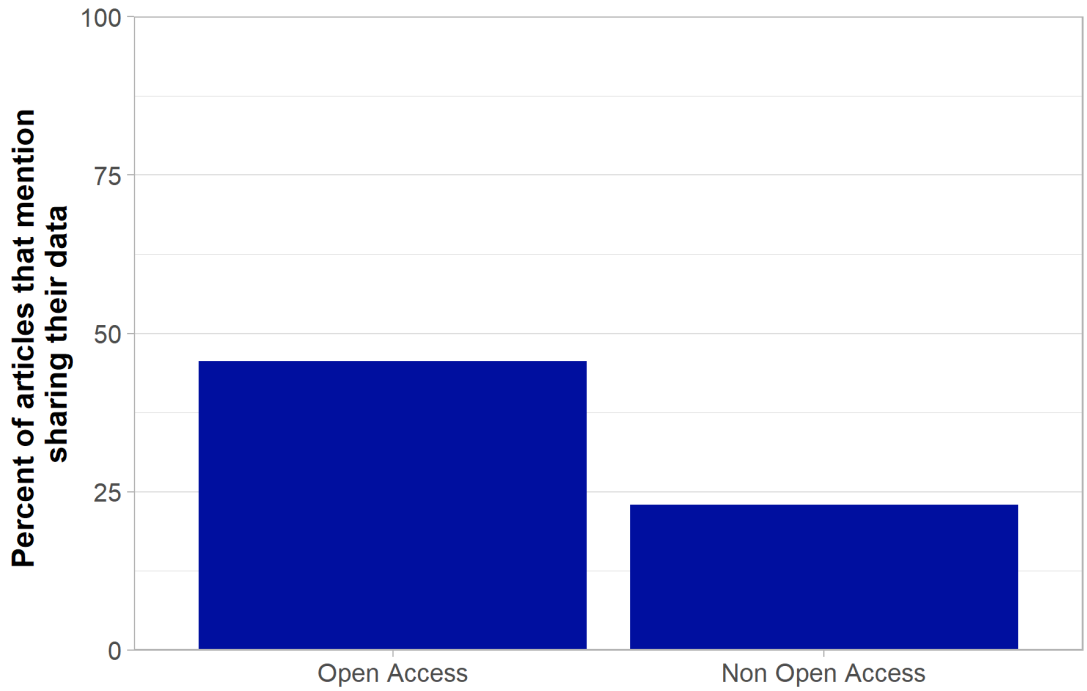

Figure 6: Percent of articles that mention sharing their data in open access versus non-open access articles.

Discussion

This single institution analysis begins to highlight the recent data sharing practices of biomedical researchers. At our institution, most articles that include original research data do not clearly state how to access that data, or if the data is accessible at all. When data is shared, it is often shared in proprietary formats, such as Microsoft Excel files. The location where reused data can be accessed was rarely shared, which is understandable, as in the case of our institution this was mainly patient health data.

These results indicate that data sharing mandates do seem to have an impact. The NIH made genomic data sharing mandatory in 2014, and genomic data was by far the most common data shared by UMass Chan researchers. Studies prior to 2014 also note a slow rise in genomic data sharing (Piwowar 2011; Hampton et al. 2013), signaling the effects of those mandates.

Our findings are similar to those of other recent research, in that data is most often listed as available upon request, followed by available in a repository (Henderson 2022). Some of the same repositories were also found to be the most popular: GEO, Sequence Read Archive, Protein Data Bank, and Figshare (Borghi 2021), which is unsurprising given that genomic data was the type of data shared most often. A previous study by Read et al. (2015), had identified ClinicalTrials.gov as one of the most popular data repositories mentioned in NIH-funded articles but ClinicalTrials.gov was not mentioned as a data repository for any of

our articles. While the repositories have changed, the amount of "invisible data" has not shifted far from the 88% of articles identified by Read et al. in 2015.

Open access availability was correlated with data availability which matches a similar finding by Piwowar (2011). This is potentially due to publisher mandates: open access journals may be more likely to require data sharing. PLOS journals, for example, "require authors to make all data necessary to replicate their study's findings publicly available without restriction at the time of publication" (PLOS ONE 2019). Previous studies have noted that the presence of a journal data sharing mandate improving the rate of finding data over journals which have no data sharing mandate (Vines et al. 2013). Future research could look in more detail at the types of open access publishing and the mention of data availability within the article.

There are, of course, limitations to this research. Our sample is from a single, biomedical institution during a single year. Our sample year is from prior to the COVID-19 pandemic, when the importance of open access to research became very clear to many (Harrison et al. 2021; Ewers, Ioannidis, and Plesnila 2021). The FAIR principles have also become more well-known with 66% of respondents from the 2021 State of Open Data Survey being familiar with them (Digital Science et al. 2021). This research could be rerun in a few years, to gauge the effects of the pandemic and the NIH Data Management and Sharing Policy. Broadening the sample to include other medical schools and biomedical research universities is also important to gain a clearer understanding of current data sharing practices and how they are evolving. Standardizing a methodology for collecting this type of information would allow for meta-analysis across institutions.

Another limitation is that our search was limited to PubMed. While most biomedical literature would be captured by a PubMed search, particularly biomedical research from an institution primarily funded by the NIH, to gain a full picture of the publishing and data sharing practices of the researchers at our institution, other databases should also be included, as well as preprint repositories. A potential area for future research would be to compare data sharing practices of those who publish on preprint servers versus those who do not.

Beyond understanding the current status of data sharing at UMass Chan, our results are important in that they help to inform future data services at our and other institutions. For example, 54% of data sets did not have an associated identifier. This would be an important topic to bring up in future data management workshops. Other potential areas to improve services would include increasing resources on best practices for sharing non-genomic data, the importance of data availability statements, and data citation. If data reuse is to become easier to track and for researchers to get credit, we need to encourage citing data in references or promote machine learning tools such as those created by the MICA project (Lafia et al. 2021).

Conclusions

In conclusion, while some researchers at the University of Massachusetts Chan Medical School have embraced data sharing, particularly genomic data sharing, we expect there will be more data shared in the coming years with the implementation of the new NIH Data Management and Sharing Policy. As journal and funder policies for more open and accessible data become mandates rather than suggestions, data librarians have a chance to become essential partners in the training and support of researchers in their data sharing needs.

Acknowledgements

Thank you to Lisa Palmer and Rebecca Reznik-Zellen for designing and completing the pilot study which informed this project.

The content of this article is based on the Poster presentation at RDAP Summit 2022, available from Open Science Framework: https://osf.io/tx9a7.

Data Availability

The data and code are available in eScholarship@UMassChan:

Grynoch, Tess and Kimberly MacKenzie. 2022. "Data and Code from 'Show Me the Data! Data Sharing Practices Demonstrated in Published Research at the University of Massachusetts Chan Medical School.'" [Data set and code]. eScholarship@UMassChan. https://doi.org/10.13028/BPQ6-HF10.

Competing Interests

The authors declare that they have no competing interests.

References

Borghi, John. 2021. "Identifying the Who, What, and (Sometimes) Where of Research Data Sharing at an Academic Institution." May 5. https://doi.org/10.5281/zenodo.4739773.

Digital Science, Briony Fane, Paul Ayris, Mark Hahnel, Iain Hrynaszkiewicz, Grace Baynes, and Emily Farrell. 2019. "The State of Open Data Report 2019." https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_Report_2019/9980783/2.

Digital Science, Natasha Simons, Greg Goodey, Megan Hardeman, Connie Clare, Sara Gonzales, Damon Strange, et al. 2021. "The State of Open Data 2021." https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2021/17061347/1.

Ewers, Michael, John P.A. Ioannidis, and Nikolaus Plesnila. 2021. "Access to Data from Clinical Trials in the COVID-19 Crisis: Open, Flexible, and Time-Sensitive." *Journal of Clinical Epidemiology* 130(February): 143–146. https://doi.org/10.1016/j.jclinepi.2020.10.008.

Grynoch, Tess and Kimberly MacKenzie. 2022. "Data and Code from 'Show Me the Data! Data Sharing Practices Demonstrated in Published Research at the University of Massachusetts Chan Medical School.'" [Data set and code]. eScholarship@UMassChan. https://doi.org/10.13028/BPQ6-HF10.

Hampton, Stephanie E., Carly A. Strasser, Joshua J. Tewksbury, Wendy K. Gram, Amber E. Budden, Archer L. Batcheller, Clifford S. Duke, and John H. Porter. 2013. "Big Data and the Future of Ecology." *Frontiers in Ecology and the Environment* 11(3): 156–162. https://doi.org/10.1890/120103.

Harris, Paul A., Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. 2009. "Research Electronic Data Capture (REDCap)—A Metadata-Driven Methodology and Workflow Process for Providing Translational Research Informatics Support." *Journal of Biomedical Informatics* 42(2): 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

Harrison, Peter W., Rodrigo Lopez, Nadim Rahman, Stefan Gutnick Allen, Raheela Aslam, Nicola Buso, Carla Cummins, et al. 2021. "The COVID-19 Data Portal: Accelerating SARS-CoV-2 and COVID-19 Research through Rapid Open Access Data Sharing." *Nucleic Acids Research* 49(W1): W619–W623. https://doi.org/10.1093/nar/gkab417.

Henderson, Margaret. 2022. "How Are Researchers at San Diego State University Sharing Their Data: Helping to Ensure Compliance." Presented at the Medical Library Association Annual Meeting 2022, May 5. https://www.eventscribe.net/2022/mla22/agenda.asp?startdate=5/5/2022&enddate=5/5/2022.

Lafia, Sara, Jeong-Woo Ko, Elizabeth Moss, Jinseok Kim, Andrea Thomer, and Libby Hemphill. 2021. "Detecting Informal Data References in Academic Literature." Retrieved from Deep Blue Documents. https://doi.org/10.7302/1671.

National Institutes of Health. 2014. "NIH Genomic Data Sharing Policy." NIH. August 27, 2014. https://grants.nih.gov/grants/guide/notice-files/not-od-14-124.html.

———. 2020a. "Budget." NIH. June 29, 2020. https://www.nih.gov/about-nih/what-we-do/budget.

———. 2020b. "Final NIH Policy for Data Management and Sharing." NIH. October 29, 2020. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.

———. 2022a. "Data Management and Sharing Overview." NIH Scientific Data Sharing. April 5, 2022. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/data-management-and-sharing-policy-overview#before.

———. 2022b. "Data Submission and Release Expectations." NIH Scientific Data Sharing. April 5, 2022. https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/data-submission-and-release-expectations.

———. 2022c. "Repositories for Sharing Scientific Data." NIH Scientific Data Sharing. April 5, 2022. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data.

Piwowar, Heather A. 2011. "Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data." Edited by Cameron Neylon. *PLoS ONE* 6(7): e18657. https://doi.org/10.1371/journal.pone.0018657.

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. 2018. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles." *PeerJ* 6(February): e4375. https://doi.org/10.7717/peerj.4375.

PLOS ONE. 2019. "Data Availability." PLOS. December 5, 2019. https://journals.plos.org/plosone/s/data-availability.

Read, Kevin B., Jerry R. Sheehan, Michael F. Huerta, Lou S. Knecht, James G. Mork, Betsy L. Humphreys, and NIH Big Data Annotator Group. 2015. "Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study." *PLoS ONE* 10(7): e0132735. https://doi.org/10.1371/journal.pone.0132735.

Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data Sharing by Scientists: Practices and Perceptions." Edited by Cameron Neylon. *PLoS ONE* 6(6): e21101. https://doi.org/10.1371/journal.pone.0021101.

UMass Medical School Communications. 2021. "UMMS Soars in NIH Funding, Blue Ridge Institute Rankings." *UMass Chan News* (blog). February 24, 2021. https://www.umassmed.edu/news/news-archives/2021/02/umms-soars-in-nih-funding-blue-ridge-institute-rankings.

Vines, Timothy H., Rose L. Andrew, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Nolan C. Kane, Jean-Sébastien Moore, et al. 2013. "Mandated Data Archiving Greatly Improves Access to Research Data." *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 27(4): 1304–1308. https://doi.org/10.1096/fj.12-218164.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(1): 160018. https://doi.org/10.1038/sdata.2016.18.