ISSN 2161-3974

JeSLIB 2021; 10(4): e1216
https://doi.org/10.7191/jeslib.2021.1216

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

Commentary

Preparing a Data Archive or Repository for Changing Research Data and Materials Retention Policies

Jonathan Bohan and Lynda Kellam

Cornell University, Ithaca, NY, USA

Abstract

Archival expectations and requirements for researchers' data and code are changing rapidly, both among publishers and institutions, in response to what has been referred to as a "reproducibility crisis." In an effort to address this crisis, a number of publishers have added requirements or recommendations to increase the availability of supporting information behind the research, and academic institutions have followed. Librarians should focus on ways to make it easier for researchers to effectively share their data and code with reproducibility in mind. At the Cornell Center for Social Sciences, we have instituted a Results Reproduction Service (R-Squared) for Cornell researchers. Part of this service includes archiving the R-Squared package in our CoreTrustSeal certified Data and Reproduction Archive, which has been rebuilt to accommodate both the unique requirements of those packages and the traditional role of our data archive. Librarians need to consider roles that archives and institutional repositories can play in supporting researchers with reproducibility initiatives. Our commentary closes with some suggestions for more information and training.

Correspondence: Jonathan Bohan: jb2287@cornell.edu

Received: June 2, 2021 **Accepted**: October 18, 2021 **Published**: November 10, 2021

Disclosures: The authors report no conflict of interest. The substance of this article is based upon a lightning talk presentation at RDAP Summit 2021. Additional information at end of article.

Archival expectations and requirements for researchers' data and code are changing rapidly among both publishers and institutions. As recently as 2016, Cynthia R.H. Vitale pointed out that "many libraries are beginning to evaluate what role they may play in improving the reproducibility of the research conducted on their campuses," but that this evaluation was "still mostly in the exploratory phase" (Vitale 2016, 38). A review of social science articles published between 2014 and 2017 showed that fewer than 20% of articles had a data availability statement, and just over 10% had a materials availability statement (Hardwicke et al. 2020, 5-6). Moreover, a search of Google Scholar shows more than 4,000 articles discussing the "reproducibility crisis" just since 2017.

In an effort to address this crisis, a number of publishers have added requirements or recommendations to increase the availability of supporting information behind the research. For example, the editorial policy for *Science* journals states that "All data used in the analysis must be available to any researcher for purposes of reproducing or extending the analysis," and that "all computer code central to the findings being reported should be available to readers to ensure reproducibility" (Science, n.d.). The American Geophysical Union "encourages authors to identify and archive their data in approved data centers," while defining data to include code and computer software used to generate results (American Geophysical Union 2016). Finally, the *American Journal of Political Science* states that authors "must provide materials that are sufficient to enable interested researchers to verify all of the analytic results" (American Journal of Political Science, n.d.).

In academia, a key challenge is finding a common terminology. Data policies are coalescing around journal requirements; however, related but not identical terms are often used interchangeably. For example, reproducibility is generally meant to indicate use of the same data to get the same results as the researcher; whereas replication means collecting new data to achieve the same results (National Academies of Sciences, Engineering, and Medicine 2019). As Sayre and Riegelman (2018, 3) point out, terminology is often misused, even within the same organization across various disciplines and in different situations. Furthermore, the Curating for Reproducibility Consortium (CURE) lists "Lack of clarity on standards for computational reproducibility" as one of the top challenges in performing reproduction work (Peer et al. 2021, 2).

Moreover, this lack of clarity can be seen in the varying language used by different institutions within university policy documents. At Cornell, an interim research data retention policy was instituted last year which states that "Reproducibility is essential to the advancement of science and requires access to relevant research data, materials, documents, protocols, methods, and procedures" (Cornell

University Policy Office 2020). The University of Pittsburgh, in their policy, says that "records should include sufficient detail to permit examination for the purpose of replicating the research" (University of Pittsburgh 2009). Finally, University of Mississippi Medical Center's policy defines research data to include "laboratory notebooks, as well as any other records that are necessary for the reconstruction and evaluation of reported results of research and the events and processes leading to those results" (University of Mississippi Medical Center 2015). Institutions are creating policies to address the proliferation of data and the reproducibility crisis, but the lack of a consistent vocabulary can create some confusion.

What does this mean for data librarians and archivists? These diverse requirements have repercussions for libraries as they require retaining different materials to fulfill the charge. For example, reproduction requires the original data and code, whereas replication would require access to more detailed information about the conduct of the original research including what variables were collected and the methods used for collection. In addition to the lack of clarity around language, researchers find that sharing research materials effectively can be difficult for a number of reasons including differing data formats and communicating context. Feinberg et. al explain that "merely uploading one's data to a public repository seldom provides sufficient context to enable others to understand and reuse it" (Feinberg et al. 2020, 35-2, 35-5). Libraries can and do fill the role of clarifying context, but it is important for librarians to understand the process behind reproducibility and ensure the repository is responsive to that process.

Making it easier for researchers to effectively share their data and code with reproducibility in mind should be a priority. In addition, libraries should convince researchers that placing their materials in a trusted repository is the best method to meet the requirements of publishers and institutions. Requirements to make research materials available can be technically met through uploading to sites like Github or Figshare or a researcher's own website, a process that requires less effort and time than depositing in an institutional archive or repository. To counter this, data archivists and librarians need to make clear the value they can add to materials deposited into a trustworthy repository.

In many cases changes to the archive may be required to provide added value. Most data archives have not been designed with code archiving in mind but have focused on data files, which are normally not connected to a specific manuscript. The archival package needed to conduct a reproduction study may contain dozens to hundreds of files which need to be run in a precise order using specific software. In addition, code is not data; it requires extensive documentation explaining the processes used to create results, but it also needs the data itself to be useful. Storing code with data can be wasteful of valuable storage space when data may be used for multiple publications. Librarians need to ask whether it makes sense to store the data together with the code or should each have separate but linked catalog records? Also for consideration, materials are related to a specific publication and can be viewed as part of that publication. For the archive, this can mean linking to that publication, external to the archive; it can also mean the materials may need to be embargoed until publication or deposited and made available in a much faster turn-around time than traditional data is deposited and made available. Data librarians need to ensure that these considerations have been addressed within the archive and that storage and retrieval of the materials is done in such a way as to make it obvious to the person performing the reproduction work the exact methods by which the researcher performed the analysis.

At the Cornell Center for Social Sciences, we have instituted a Results Reproduction Service for Cornell researchers, named “R-squared” (Cornell Center for Social Sciences, n.d.). Our consultants work with the researchers to ensure that their code and data produce the results in their publication or publications. Consultants reproduce output precisely, offer suggested code edits to facilitate reproducibility, and assist in creating a ReadMe file that provides detailed instructions for running the code.

As part of this service, we offer to archive the entire package in our CoreTrustSeal certified Data and Reproduction Archive. This package of data and code is put into a zip archive with a ReadMe file containing detailed instructions for running the code, including information on the exact version of the software used. This allows researchers to reproduce the published results; the archival record for this “R-squared” package is stored within a separate database from our data archive but is findable through the same search functionality. We provide a suggested citation, a persistent identifier for the materials themselves, links to the researchers’ ORCID or ResearchGate profiles, and a special citation for the “reference article.” We can choose to archive the data within the Zip archive, or in a separate record within the data archive if we believe the data may be used for multiple publications. If the data are archived separately, we provide precise instructions in the original R-squared package on how to integrate the data into the reproduction materials.

At CCSS we re-built our Data & Reproduction Archive to accommodate both the unique requirements of R-squared packages and the traditional role of our data archive. We found that creating our own repository would meet the needs of our R-Squared services and researchers, as well as provide the flexibility needed for rapidly changing requirements and terminology. As such, our solution works for both our traditional role as a trusted data repository and the transparency needs of our researchers.

Nevertheless, the data ecosystem is a changing landscape. The challenge for any library or data repository is to remain responsive to changes in technology and user needs while looking for possibilities to improve. When it comes to reproduction materials, as Kapiszewski and Karcher put it, "[e]ven at large research institutions . . . libraries often lack the information technology and subject-specific capabilities to provide curation, preservation, and dissemination guidance and services on par with domain repositories" (Kapiszewski and Karcher 2020, 207-208). Librarians who are interested can seek training in reproduction methods, such as workshops offered by the Curating for Reproducibility (CURE) consortium (Peer 2019), or libraries can consider hiring reproducibility specialists or training students to assist in the process. As always, academic librarians and libraries should be aware of the needs of researchers and be responsive to their needs and the needs of the changing data ecosystem.

Acknowledgments

The authors would like to acknowledge the contributions of CCSS staff, especially Florio Arguillas, to our understanding of quantitative reproduction methods and Brandon Kowalski to the design of CCSS's Data & Reproduction Archive. We would also like to thank Kathleen Weldon of The Roper Center for Public Opinion Research for reviewing the draft of this manuscript. Lastly, we would like to thank William Block, former director of the Cornell Institute for Social and Economic Research.

Disclosures

The content of this article is based upon a lightning talk presentation at RDAP Summit 2021 titled "Preparing a Data Archive or Repository for Changing Research Data and Materials Retention Policies" available at https://osf.io/ug5ty.

References

American Geophysical Union. 2016. "Data Policy." AGU Policy: Data. Accessed April 12, 2021. https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Policies/Data-policy

American Journal of Political Science. n.d. "AJPS Verification Policy." Accessed April 12, 2021. https://ajps.org/ajps-verification-policy

Cornell Center for Social Sciences. n.d. "Results Reproduction (R-Squared)." Accessed April 12, 2021. https://ciser.cornell.edu/research/results-reproduction-r-squared-service

Cornell University Policy Office. 2020. "Interim Policy 4.20 - Research Data Retention." Accessed April 12, 2021. https://www.dfa.cornell.edu/policy/policies/research-data-retention

Curating for Reproducibility. n.d. “CURE Mission.” Accessed May 25, 2021. https://cure.web.unc.edu

Feinberg, Melanie, Will Sutherland, Sarah Beth Nelson, Mohammad Hossein Jarrahi, and Arcot Rajasekar. 2020. “The New Reality of Reproducibility: The Role of Data Work in Scientific Research.” *Proceedings of the ACM on Human-Computer Interaction* 4(CSCW1): Article 035. https://doi.org/10.1145/3392840

Hardwicke, Tom E., Joshua D. Wallach, Mallory C. Kidwell, Theiss Bendixen, Sophia Crüwell and John P. A Ioannidis. 2020. “An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017).” *Royal Society Open Science* 7(2). http://doi.org/10.1098/rsos.190806

Kapiszewski, Diana and Sebastian Karcher. 2020. “Making Research Data Accessible.” in *The Production of Knowledge: Enhancing Progress in Social Science*, edited by Colin Elman, John Gerring, and James Mahoney: 197-220. Strategies for Social Inquiry. Cambridge: Cambridge University Press, 2020. http://doi.org/10.1017/9781108762519.008

National Academies of Sciences, Engineering, and Medicine. 2019. “New report examines reproducibility and replicability in science.” *Phys.org* May, 7 2019. https://phys.org/news/2019-05-replicability-science.html

Peer, Limor. 2019. “Curating for FAIR and Reproducible Data and Code.” Research Data Alliance. Accessed May 27, 2021. https://www.rd-alliance.org/curating-fair-and-reproducible-data-and-code

Peer, Limor, Florio Arguillas, Tom Honeyman, Nadica Miljković, Karsten Peters-von Gehlen and CURE-FAIR WG Subgroup 3. 2021. “Challenges of Curating for Reproducible and FAIR Research Output.” https://doi.org/10.15497/RDA00063

Sayre, Franklin and Amy Riegelman. 2018. “The Reproducibility Crisis and Academic Libraries.” *College & Research Libraries* 79(1): 2-9. https://doi.org/10.5860/crl.79.1.2

Science. n.d. “Science journals: editorial policy.” Accessed April 12, 2021. https://www.sciencemag.org/authors/science-journals-editorial-policies

University of Pittsburgh. 2009. “Guidelines on Research Data Management.” Accessed April 12, 2021. http://www.provost.pitt.edu/documents/RDM_Guidelines.pdf

University of Mississippi Medical Center. 2015. “Policy on Research Data Retention.” Accessed April 12, 2021. https://www.umc.edu/Research/files/Policies%20and%20Procedures%20Files/policy-on-research-data-retention.pdf

Vitale, Cynthia R.H. 2016. “Is Research Reproducibility the New Data Management for Libraries?” *Bulletin of the Association for Information Science and Technology* 42(3): 38-41. https://doi.org/10.1002/bul2.2016.1720420313