Why and How We Share Reproducible Research

Limor Peer

Research institutions have an interest in verifying the reproducibility of replication packages before they are published. The main reasons for this are research institutions’ commitment to responsible research, the need to build competencies and capacity around open research, and the culture change around open research. I offer this perspective based on the experience of a data archive based at the Institution for Social and Policy Studies (ISPS) at Yale University, which has over 10 years of experience verifying and sharing reproducible research.

## ISPS Data Archive

The Institution for Social and Policy Studies (ISPS) was founded in 1968 as an interdisciplinary center to support social science and public policy research at Yale University. Since the year 2000, ISPS has pioneered and established itself as an important center for the implementation of randomized controlled trials (or “field experiments”), especially in political science where use of the method was quite uncommon at that time. About a decade later, ISPS created an archive for sharing and archiving the accumulated body of knowledge from these experiments, as well as other data-driven computational social science research. Affiliated researchers are invited to deposit data and code and associated materials used to produce original research (i.e., a “replication package”) in the ISPS Data Archive.

Since those days in the early 2010s, dramatic developments in the scholarly landscape have taken place around research data management and sharing – in terms of infrastructure, tools, and guidelines. Researchers have more choices where to deposit and share their data and code and they are increasingly required or encouraged to do so by journals and funders. The emergence and broad acceptance of the FAIR principles in the last decade propelled all stakeholders to work toward implementing better practices (e.g., data citation) and following standards (e.g., use of PID). However, there is less agreement on what other standards to prioritize and – such as, independent understandability, long-term reusability, reproducibility – or on who bears responsibility to uphold them. ISPS set out to take a broad view of open research that acknowledges all relevant standards.

### Vision and standards

From the archive’s early days, ISPS approach was that it has a responsibility to assist researchers who wish to disseminate and archive research products that support published scientific claims. ISPS assistance, it was determined, includes a review of the replication package that will result in confirmation that the materials indeed support the reported findings.

“ISPS believes that it has both responsibility and expertise to assist researchers who wish to disseminate and archive research products that support published scientific claims and has created a process to ensure computational reproducibility.” (<https://isps.yale.edu/research/data/approach>)

This orientation to responsible data and code sharing stems from strongly held principles. First, a set of values related to research ethics: rigor, transparency, and integrity. The idea is that rigorous research practices must extend to best practices in dissemination and archival. Second, the values of stewardship, curation, and preservation. These concepts represent a commitment to the longevity of these materials. In addition to compliance with all legal and ethical limitations, responsible data and code sharing also demands that other standards are met, for example, that materials are usable, reproducible, and independently understandable. Upholding these standards confers credibility to the research and aligns with the scientific ethos which elevates the ability to extend and build upon previous findings. In short, the ISPS vision is to enable continued access and independent reuse of the research compendium for the long term and to ensure that the quality of the objects meets community standards for FAIR and for long-term archival preservation.

### A well-tended garden

ISPS chose to strongly recommend but not require its affiliated researchers to engage with the review process or deposit in the ISPS Data Archive prior to sharing data and code. Instead, ISPS focuses on the practical benefit that responsible data and code sharing provides researchers: It offers an opportunity for researchers to have professionals review materials before they are shared with the scientific community.

The review function of the ISPS Data Archive is implemented by means of a “push” or a “pull.” Researchers can request a review prior to submission to a journal or during the journal review process (“push”), or ISPS will obtain copies of a replication package made available elsewhere or otherwise request a replication package from the researcher and perform the review (“pull”). In all cases, ISPS review will result in the deposit of the replication package in the ISPS Data Archive and its publication on the ISPS website.

The curation team communicates with researchers on any issues that surface during the review. Replication packages are published when computational reproducibility is verified. In cases where full verification is not achievable or feasible, a curator README is published as part of the replication package (see WORKING PAPER). This is a small collection for a designated community. As of June 2023, the archive holds a collection of over 120 verified replication packages.

### Framework and workflow

ISPS’ focus on responsible data and code sharing led it to develop policies, tools, and workflows in support of that goal. ISPS developed in-house expertise and stood up a process to review research materials underlying research claims, including to ensure the computational reproducibility of replication packages (see [<https://isps.yale.edu/research/data/approach>](https://isps.yale.edu/research/data/approach)).

The process is based on the Data Quality Review framework, which prescribes actions to curate data and code and to review the computational reproducibility of the materials (Peer et al., 2014). ISPS developed software to facilitate the curation and review workflow, the Yale Application for Research Data (YARD).

### Institutional support

The ISPS Data Archive officially launched in 2011 (Peer & Green, 2012; Peer, 2022). In addition to ISPS commitment of resources, the archive benefitted from the support of university partners including the Office of Data and Assent Infrastructure (ODAI), which sponsored the initial ISPS pilot; the Yale University Library, which consults on solutions and service around data management, curation, and preservation; Yale University IT and Central IT, which took on support of the YARD technology; and the recently-established Data-Intensive Social Science Center which is co-sponsoring the archive as part of its research support services portfolio.

## Thoughts on reproducibility in the context of open research

The original question posed to the panel was, why can or should research institutions publish replication packages? I answer this question with one of my own: What is required of research institutions in the age of open research? My position is that whether research institutions choose to publish replication packages or not, they have an interest in verifying the reproducibility of replication packages before they are published. This is for several reasons.

First, institutions have an interest to produce research that adheres to the highest standards at every stage of the research lifecycle. Institutions committed to responsible research – whether entire universities or centers within – must drive the development of socio-technical infrastructure that supports the values of rigor, transparency, and integrity. Academic institutions tend to address these issues early in the research lifecycle, for example, via Institutional Review Boards. However, commitment to stewardship and preservation, especially in the age of open research, calls for some level of internal review in later stages of the research lifecycle as well. If executed well, a bit of useful friction can reduce cost in the system overall, for example, by minimizing minor annoyances to others attempting to use the materials.

“No one is immune from making mistakes. In research, mistakes might include analyzing raw data instead of cleaned data, reversing variable labels, transcribing information incorrectly, or inadvertently saving over a file. The consequences of these kinds of mistakes can range from minor annoyances like wasted time and resources to major issues such as retraction of an article…” (Strand, 2023)

Second, it follows that institutions need to develop competencies around open research or open science. In the Unites States, as elsewhere, calls for public access to scientific research, including data and code, are intensifying (for example, 2013 OSTP memo, 2022 Nelson memo). Open research is understood as better science. These calls are motivated by the goal of getting more value out of the public investment in science and making research more reproducible and transparent. At the same time, there is recognition that technology alone cannot meet the challenge of open research (Ayris, 2016). Institutions can develop in-house capacity by creating services in this area, training researchers and staff on various aspects of open research that can enhance reproducibility (e.g., open source software, coding skills, version control), and providing discipline-specific research infrastructure (e.g., computer clusters, large data storage). This can include competencies around reproducibility verification (as well as other functions such as data curators and stewards).

“In particular, teaching students to work reproducibly enables easier and deeper evaluation of their work; having them reproduce parts of analyses by others allows them to learn skills like exploratory data analysis that are commonly practiced but not yet systematically taught; and training them to work reproducibly will make their post-graduation work more reliable.” (Donoho, 2017)

Institutions better positioned to deliver open research will be more competitive. Benefits of a robust institution-based open research capacity include bridging the chasm between e-infrastructure providers and scientific domain specialists (Ayris, 2016) and building institutional memory for research projects (Nolan, 2023).

In conclusion, institutions have an interest in adapting, if not leading, the culture change around open research. Publishers have made advances in this area and are increasingly subjecting replication packages to review and verification (see relevant chapter in this book). However, as the National Academies of Science, Engineering, and Medicine point out, “this process has not been adopted at most journals because it requires a major commitment of resources.” (NASEM 2023, p.190) Institutions are well-positioned to bridge gaps – or act as a pressure point –between researchers and publishers as well as funders to help bring turn vision to reality.

## References

Ayris, P., Berthou, J-V., Bruce, R., Lindstaedt, S., Monreale, A., Mons, B., Murayama, Y., SödergSrd, C., Tochtermann, K., Wilkinson, R. (2016). Realising the European Open Science Cloud. European Commission. Retrieved from https://ec.europa.eu/research/openscience/pdf/realising\_the\_european\_open\_science\_cloud\_2016.pdf

Donoho, D. (2017). 50 Years of Data Science, Journal of Computational and Graphical Statistics, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734.

National Academies of Sciences, Engineering, and Medicine. 2023. Behavioral Economics: Policy Impact and Future Directions. Washington, DC: The National Academies Press. DOI: 10.17226/26874.

Nolan, R. (2023). Building Institutional Memory for Research Projects – Why education is key to long-term change. LSE Impact Blog. Retrieved from: https://blogs.lse.ac.uk/impactofsocialsciences/2023/02/21/building-institutional-memory-for-research-projects-why-education-is-key-to-long-term-change/.

Peer, L. (2022). Ten Years of Sharing Reproducible Research. ResearchDataQ (Association of College and Research Libraries). Retrieved from https://researchdataq.org/editorials/ten-years-of-sharing-reproducible-research/.

Peer, L., Green A. (2012). Building an Open Data Repository for a Specialized Research Community: Process, Challenges, and Lessons. International Journal of Digital Curation. 7(1): 151-162. DOI: 10.2218/ijdc.v7i1.222

Peer, L., Green, A., Stephenson, E. (2014). Committing to Data Quality Review. International Journal of Digital Curation. 9(1): 263-291. DOI: 10.2218/ijdc.v9i1.317.

Strand, J. F. (2023). Error tight: Exercises for lab groups to prevent research mistakes. *Psychological Methods.* Advance online publication. [DOI: 10.1037/met0000547](https://psycnet.apa.org/doi/10.1037/met0000547).

Working paper (in progress).