Collaborative Librarianship

Volume 9 | Issue 2 Article 4

7-11-2017

Reproducibility Librarianship

Vicky Steeves
New York University, vicky.steeves@nyu.edu

Reproducibility Librarianship

Cover Page Footnote
Supplementary Materials Accompanying data and scripts can be found at this Open Science Framework project: https://osf.io/nzx2e/. Acknowledgements I would like to acknowledge Dr. Juliana Freire, the Principal Investigator of ReproZip and Director of Graduate Studies at the CDS, and Scott Collard, the Head of Specialized Research Services in the Division of Libraries, for their unwavering support. I'd also like to thank Remi Rampin, Nicholas Wolf, Jeffrey Spies, and Franklin Sayres for their help in looking over this paper.

Steeves: Reproducibility Librarianship

From the Field

Reproducibility Librarianship

Vicky Steeves (vicky.steeves@nyu.edu)
Librarian for Research Data Management & Reproducibility, New York University

Abstract

Over the past few years, research reproducibility has been increasingly highlighted as a multifaceted challenge across many disciplines. There are socio-cultural obstacles as well as a constantly changing technical landscape that make replicating and reproducing research extremely difficult. For example, the prioritization of citation counts and journal prestige has undermined incentives to make research reproducible. Technically, researchers face challenges in reproducing research across different operating systems and different versions of software.

While libraries have been building support around research data management and digital scholarship, reproducibility is an emerging area that has yet to be systematically addressed. In response, New York University created the position of Librarian for Research Data Management and Reproducibility (RDM & R), a dual appointment between the Center for Data Science (CDS) and the Division of Libraries. This report will outline the role of the RDM & R librarian, with special focus on the collaboration between the CDS and Libraries to bring reproducible research practices into the norm.

Keywords: reproducibility, data management, data librarianship

Introduction

Spurred by the Center for Open Science's systematic examination of the state of reproducibility in psychology[1], open research and reproducibility has received considerable attention in both academia and the popular media. This scrutiny has resulted in increased discourse around reproducibility in a number of disciplines, including physical and life sciences, digital humanities, computational science, and medicine. Advocates for openness and reproducibility are lobbying for changes in methodological reporting (e.g., pre-registration of studies), more transparent analyses, improved data management, sharing of research materials, a transformation of the traditional publishing model, and reform of the promotion and tenure process.

At the center of these developments is the idea that reproducibility is a core property of research: it is not only essential for verification and authentication of results, but also for driving a field forward. If a work is reproducible, others in the field can easily build upon it. While it is easy to make materials available, given the proliferation of repositories that support diverse types of research output, reproducibility still remains an elusive target for many. For instance, the use of proprietary file formats and analysis software limit usability and reproducibility.

Anderson, Martinson, and DeVries[2] found that researchers' endorsement of scientific ideals (e.g., openness) and their behaviors don't match. Most of the surveyed scientists subscribed to the values, but did not always practice them. These

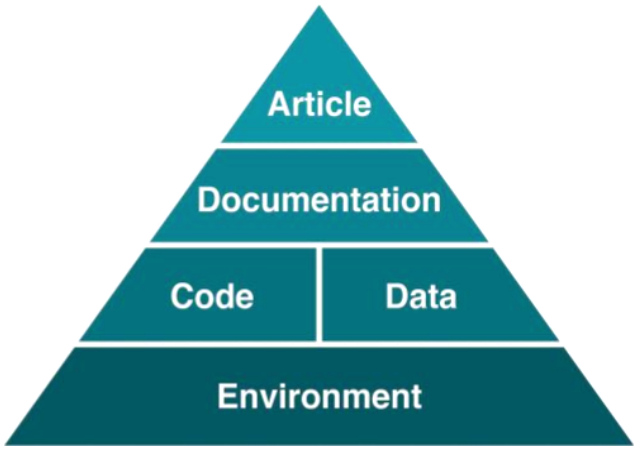
scientists even perceived a greater incongruence in their peers. It does not help that the term reproducibility—while certainly an ideal—is used inconsistently. Language around reproducibility and computational/methodological concepts of reproducibility vary across research domains. Stodden et al.[3] define the spectrum of reproducibility as follows:

> **Reviewable Research.** The descriptions of the research methods can be independently assessed and the results judged credible. (This includes both traditional peer review and community review, and does not necessarily imply reproducibility.)
>
> **Replicable Research.** Tools are made available that would allow one to duplicate the results of the research, for example by running the authors' code to produce the plots shown in the publication. (Here tools might be limited in scope, e.g., only essential data or executables, and might only be made available to referees or only upon request.)
>
> **Confirmable Research.** The main conclusions of the research can be attained independently without the use of software provided by the author. (But using the complete description of algorithms and methodology provided in the publication and any supplementary materials.)
>
> **Auditable Research.** Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.
>
> **Open or Reproducible Research.** Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

Libraries have provided support across the spectrum of reproducibility. For reviewable research, librarians are often asked to do peer review, both internal and for publications. For auditable research, librarians are engaged in designing, building, and maintaining research infrastructure that ensures integrity and authenticity such as repositories and digital archives. To a degree, information professionals even support replicable research by providing embargo features and access restrictions in research infrastructure (e.g., refereeing access on behalf of the researcher).

An increasing reliance on digital tools has created new challenges. Releasing code and data are key to open research, but not necessarily enough for reproducibility. This is where the concept of computational reproducibility becomes important. Researchers used to capture their research environments with drawings; now, researchers and the librarians who work with them must capture digital environments for reproducibility (see Fig 1 below).

Figure 1. Reproducibility Pyramid. Image courtesy of Andrew Rarig (NYU)



Preserving digital environments is difficult to do. Tracking these dependencies is challenging – there are many layers of hardware and software that the average user has no skill or time to examine.[4] In Victoria Stodden's 2010 survey of the machine learning community, she learned that most authors 'claim that they do not have time to document and clean up the code'[5]. Gronenschild, et. al went into more depth with computational reproducibility, and discussed how the results of data analyses in neuroscience performed with the same application differed based on the operating system, workstation type, and software version[6]. This represents an area of reproducibility work yet to be undertaken systematically by libraries. A few systems like Yale University Library's "emulation as a service" or Carnegie Mellon University's Olive Archive offer legacy base operating system access to users that could in time start to address computational reproducibility as analysis software is added to their collections.

With more university libraries than not equipped with a data services team (also called data management, research data services, or statistical services), their involvement with researchers data management has grown exponentially[7]. The proffered support extends to a number of activities in the research data lifecycle: data management, open research, reporting guidelines, pre-registration, and digital scholarship services. Libraries have reliably and steadily responded to changes in patron needs in the past, developing new technologies and skill sets to address the altering research landscape. Now it is time for information professionals across libraries to respond to this current challenge: reproducibility.

Background

The Librarian for Research Data Management and Reproducibility (RDM & R) is a dual appointment between the Division of Libraries and Center for Data Science (CDS) at New York University. She works directly to support the Moore/Sloan Data Science Environment at the CDS and the Data Services department in the Libraries. The Libraries were seeking to hire expertise in research data management, and the CDS needed someone devoted to outreach and education around reproducibility. The introduction from the job description reads:

New York University Libraries and the New York University Center for Data Sciences seek an information professional with a background in the sciences and/or computer science to develop a set of wide-ranging programs in support of the Moore-Sloan Data Science Environment partnership. [...] The position serves as a Libraries team member on several working groups and is specifically focused on carrying out the activities of the Reproducibility and Open Science Working Group. The position is based in the Libraries and reports jointly to the Libraries Head of Data Services and the Chair of the Reproducibility/Open Science Working Group.[8]

By merging the two areas of expertise into one role, the University poised itself to systematically integrate reproducibility into existing library services and data management into the open science and reproducibility work of the CDS. The Librarian for RDM & R has three offices: in Bobst Library, in the CDS, and in the Visualization and Data Analytics laboratory at NYU's Brooklyn campus. These three locations offer her a chance to interface with the different patron groups for which she is to build support services.

The Librarian for RDM & R is charged with three essential activities on behalf of both the CDS and Libraries:

1. **Educational initiatives:**
   Providing instructional and consultation services in RDM to faculty and advanced students; exploring and piloting base-line

services in curation practices and techniques; and advising researchers on how to meet the data management and open data requirements of publishers and federal funding agencies[9].

2. **Outreach:**
   Establishing and maintaining an active program of events both for outreach and promotion, and for sharing new developments. In the course of this work, the incumbent would have the opportunity to delve deeply into the data lifecycle practices within a number of labs as case studies in scholars' practices as a means of understanding what interventions will be most valuable. [They are also responsible for] conducting ongoing assessment and monitoring of researcher needs across sciences disciplines and domains[10].

3. **Tool/Infrastructure building and support:**
   Developing a tools registry to promote sharing rather than reinventing tools, and applying sophisticated search techniques to help researchers identify "reproducibility badged" tools; be involved in efforts to design a data repository and storage infrastructure for researchers at the University; working closely with others in the libraries, the incumbent may work on developing methods and workflows for making large science data sets re-usable and library-preservable[11].

Each of these require a different skill set that must be used in tandem to achieve the goals of the position: technological literacy, teaching ability, and reference skills. Soft skills as well as technical competence are key to fulfilling the mandate of this position, similar to the way other data librarians work.

Day-to-Day Work

The Librarian for RDM & R is an inherently collaborative position; with one person responsible for such a diversity of work, it is important to develop relationships and sustainable workflows for meeting objectives. As a dual appointment managing cross-campus relationships and mastering technical terms (in many disciplines) is crucial for success. Metadata to a librarian is different from metadata to a neuroscience researcher and being able to connect to disparate communities is key to successful collaborations and viable services.

The Librarian for RDM & R works primarily on three teams. In the Libraries, she works with the Librarian for Research Data Management (RDM) as a part of the data management team within the Data Services department. This represents the first institutional service in research data management at NYU; the focus of the team is building new services. At the CDS she works with the ReproZip team, which includes one research engineer, and one doctoral candidate, and the Open Science and Reproducibility working group—a larger body dedicated to creating a culture of open scholarship and promoting reproducible research practices on campus. However, the position is defined such that the Librarian for RDM & R collaborates with a diverse group of researchers, (both within and outside NYU), engineers, supercomputing professionals, and digital humanists to promote openness and reproducible research practices within the United States and abroad.

These collaborations have been brought to bear through three fundamental functions the Librarian for RDM & R was tasked with fulfilling. They are as follows:

Educational Initiatives

By advocating for data management as the means towards achieving reproducibility, the classes and workshops offered on the topic always have content about reproducibility within them, including: a) best practices that enable greater reproducibility, b) ethics around open

scholarship, and c) resources on campus, like the Librarian for RDM & R and ReproZip. Each class within the library is co-taught alongside the Librarian for RDM.

The Librarian for RDM & R also teaches by request of faculty in sessions that are embedded within for-credit classes. These requests have emerged on the heels of extensive outreach work, outlined in the section below. These sessions were usually one class section and served as a primer to students on how to create well-managed, reproducible research. Data management is now a required session as a part of the Responsible Conduct of Research (RCR) course for the National Science Foundation and the National Institute for Health. These two types of requested classes offer a great opportunity for outreach and have resulted in greater adoption of reproducibility and data management services as well as institutionally supported tools. The Librarian for RDM & R also teaches one class out of the many RCR sessions, which are geared specifically toward students and postdocs who receive grant funding.

In the interest of building collaborations with those outside NYU who work on reproducibility, the Librarian for RDM & R has invited external speakers to give workshops on reproducible research practices. Most notably, she brought in a representative from the Center for Open Science to give a workshop on quantitative reproducibility, which opened the door for further work together, as outlined in the section below.

Having open research built into her job description, the Librarian for RDM & R has made all her scholarship, teaching, and outreach materials open source and available with a permissive license on GitHub. This has resulted in an increase in external collaborations in professional development activities and in service to professional organizations, as well as garnered contributions from others via GitHub's pull request feature that has improved the teaching materials.

Other educational resources for the larger community include reproduciblescience.org, which provides news and a resource directory for those getting started with reproducibility or those looking for resources of a specific kind. There is also a portion of this website with resources specific to the NYU community. Additionally, the Librarian for RDM & R has produced a research guide on reproducibility with more information and background on the subject.

Outreach

When first arriving at the University, the Librarian for RDM & R (with the Librarian for RDM) met with every liaison librarian to better understand the type of research that their patrons undertake and the type of data they generate. As liaison librarianship involves a focused and dedicated relationship with a subset of library patrons, the work is reliant on these relationships through two-way communication with their constituencies[12]. This was an invaluable resource, and conversations with liaison librarians set the framework for data management services within the Libraries, which blossomed into the reproducibility services offered through the Libraries and CDS. Together with the Librarian for RDM, the Librarian for RDM & R used openness as the foundation for data management services—everything used in these services is open source and available for others to use. This was made as an ethical and practical consideration: the team believed in the goals and missions of openness and wanted others to be able to use and contribute to their materials for years to come.

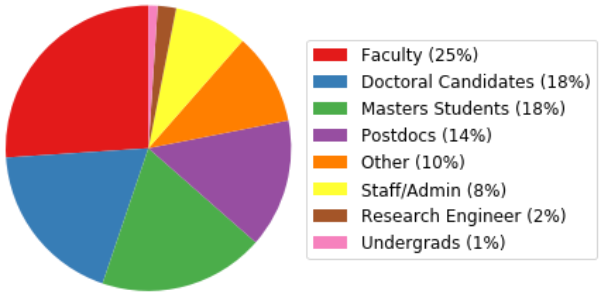
This outreach to liaison librarians was extremely fruitful and lead not only to more opportunities to develop personal research agendas[13], but also increase the number of requests and support for

reproducibility events, workshops, and classes. One specific success came on the heels of a collaborative research project with the Librarian for Life Sciences, which aimed to assess how faculty in life sciences dealt with their data and where they saw the library in their research lifecycle (publication forthcoming). The first part of this study invited faculty to participate in an anonymous online survey about their data practices. Those who agreed to be interviewed were asked more specific and targeted questions on the same topics. The study aimed to discover how faculty manage their data in order to improve library services. The survey had a 28% response rate and the request for interviews had a 16% response rate. This also led to a marked increase in requests from faculty for individual and group consultations as well as embedded classes and requests to teach RCR sessions.

Another major outreach initiative organized by the Librarian for RDM & R was the 2016 NYU Reproducibility Symposium in which members of the Moore-Sloan Data Science Environment (a collaboration between NYU, the University of California at Berkeley, and University of Washington) showcased tools and workflow to help make the reproducibility process easier, along with case studies showing how creating reproducible experiments has helped other research groups. There were 21 domains (see Table 1) represented, with about 25% of the total attendees being faculty, 18% doctoral candidates, 18% masters students, 14% postdocs, 10% "other", 8% staff/administrator, 2% research engineer, and 1% undergraduate students (see Figure 2).

To successfully disseminate reproducible research practices and the methods by which she has built services around reproducibility, the Librarian for RDM & R has engaged in multiple external collaborations to give reproducibility a wider platform. These include engaging researchers in domain-specific conferences (the European Geosciences Union conference), other librarians through professional organizations (e.g., LITA, DASPOS, PASIG, ACRL), professional development opportunities around using ReproZip and general best practices for reproducibility (see the section below).

Figure 2. Breakdown of Reproducibility Symposium attendees



Tool/Infrastructure building and support

As a core part of her work at the CDS, the Librarian for RDM & R has largely worked on ReproZip, an open source tool designed by an engineer at the CDS to help researchers overcome the technical difficulties involved in preserving and replicating research, applications, databases, software, and more (see Figure 3). The Librarian for RDM & R interfaces with users, building the development queue and contributing to documentation and other user-facing materials.

ReproZip is also core to her work in promoting reproducible research practices. It works by creating a small, self-contained package (.rpz file) by automatically identifying, tracking, and capturing all required dependencies of research processes, computation, and applications[14]. This package is easily shareable, citable, and usable by the creator and the community at large, as it is usually quite compact. Secondary users can unpack the .rpz using ReproUnzip and reproduce the work on their machine regardless of operating system. ReproUnzip's functionality is not limited to simple reproduction: it also allows

users to modify the original experiment in support of reuse and extension with very little effort.

ReproZip is extensible enough to be used for reproducibility across research domains as well as across library services. The Librarian for RDM & R has integrated ReproZip into library classes and instruction on reproducibility. Because of the simplicity of the user interface, it's been adopted by researchers across domains—from data scientists to digital humanists. Additionally, ReproZip is open-source software, which means others can contribute, modify, and extend it. Within the library, this has been useful in leveraging the tool for repository services, digital archiving, and more.

Through working on joint initiatives between Information Technology at NYU and the Libraries, the Librarian for RDM & R has been able to ingratiate these ideas of reproducibility into infrastructure planning and execution. The most notable example is the adoption of the Open Science Framework[15] at NYU, which was proposed and facilitated by the Librarian for RDM & R. The OSF is a free, open source, project and collaboration management tool built and maintained by the Center for Open Science for researchers to use throughout the research lifecycle. By enabling users easier access to a developed tool for data management, it has become easier to propagate best practices throughout the NYU community. The free, institutional offering of the OSF includes the custom nyu.edu domain name, single sign-on with University credentials, and a custom dashboard for displaying affiliated projects.

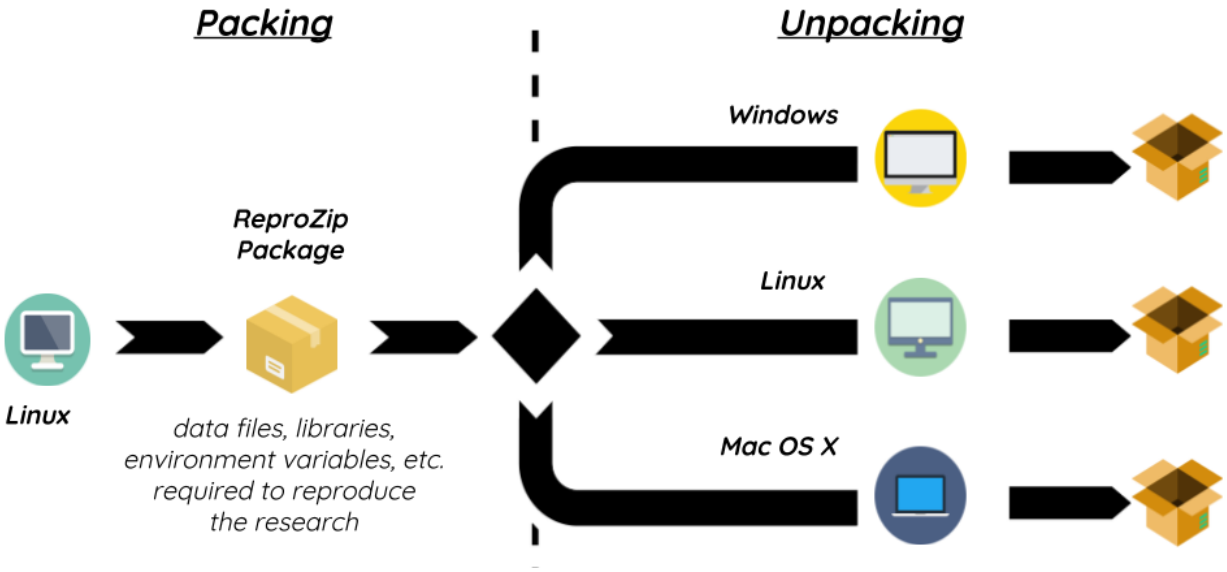
Conclusion

Providing data management and reproducibility services for a diverse and dynamic research community on campus is a demanding task that requires a distributed effort. Each service fills different gaps for researchers at various stages of their research workflow. By creating and supporting a position explicitly addressing research reproducibility and open scholarship through collaboration, New York University has begun to systematically build collaborative and sustainable services around reproducibility, extending beyond the research guide and occasional workshop to a full-blown service area. Data librarianship, while fairly recent as a sub-field, has made a huge impact on patrons in areas including data management and sharing, data management plans for grant applications, reporting guidelines, pre-registration, and scholarly communications. Acknowledging that data management services are best delivered via the library and building on this momentum, reproducibility as an integrated part of collaborative library services is the next step towards holistic research services.

Table 1. Attendees' departmental affiliation from the 2016 NYU Reproducibility Symposium

Domain	# of Attendees	Domain	# of Attendees
Psychology	11	Astrophysics	1
Data Science	7	Computational Biology	1
Applied Mathematics	4	Economics	1
Libraries	4	Education	1
Statistics	4	Financial engineering	1
Cognitive Neuroscience	3	History	1
Computer Science	3	Medical Image Processing	1
Computer Engineering	2	Nuclear Engineering	1
Human Behavior	2	Oceanography and Genomics	1
Neuroscience	2	Public Health	1
Political Science	2	**Total**	**54**

Figure 3. High-level overview of ReproZip functionality.



[1] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science* 349, no. 6251 (August 28, 2015): aac4716-aac4716, doi:10.1126/science.aac4716.

[2] Melissa S. Anderson, Brian C. Martinson, and Raymond De Vries, "Normative Dissonance in Science: Results from a National Survey of U.s. Scientists," *Journal of Empirical Research on Human Research Ethics: JERHRE* 2, no. 4 (December 2007): 3–14, doi:10.1525/jer.2007.2.4.3.

[3] Victoria Stodden, Jonathan Borwein, and David H. Bailey, "Setting the Default to Reproducible," *Computational Science Research. SIAM News* 46, no. 5 (2013): 4–6. http://stodden.net/icerm_report.pdf

[4] Ben Marwick, "How Computers Broke Science – and What We Can Do to Fix It," *The Conversation*, November 9, 2015, http://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938.

[5] Victoria C. Stodden, "The Scientific Method in Practice: Reproducibility in the Computational Sciences," 2010, https://academiccommons.columbia.edu/catalog/ac:140117.

[6] Ed H. B. M. Gronenschild et al., "The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements," ed. Satoru Hayasaka, *PLoS ONE* 7, no. 6 (June 1, 2012): e38234, doi:10.1371/journal.pone.0038234.

[7] Kathryn Crowe and Michael Crumpton, "Defining the Libraries' Role in Research: A Needs Assessment; A Case Study," 2016, https://libres.uncg.edu/ir/listing.aspx?id=19091.

[8] NYU Division of Libraries and NYU Center for Data Science, "Research Data Management and Reproducibility Librarian," accessed May 2, 2017, https://osf.io/q2bk6/.

[9] NYU Division of Libraries and NYU Center for Data Science, "Research Data Management and Reproducibility Librarian," accessed May 2, 2017, https://osf.io/q2bk6/.

[10] NYU Division of Libraries and NYU Center for Data Science, "Research Data Management

and Reproducibility Librarian," accessed May 2, 2017, https://osf.io/q2bk6/.

[11] NYU Division of Libraries and NYU Center for Data Science, "Research Data Management and Reproducibility Librarian," accessed May 2, 2017, https://osf.io/q2bk6/.

[12] Isabel D. Silver, "For Your Enrichment: Outreach Activities for Librarian Liaisons," *Reference & User Services Quarterly* 54, no. 2 (January 26, 2015): 8–14. https://journals.ala.org/index.php/rusq/article/view/2763

[13] Katherine Boss and Meredith Broussard, "Challenges Facing the Preservation of Born-Digital News Applications," 2016, http://blogs.sub.uni-hamburg.de/ifla-newsmedia/wp-content/uploads/2016/04/Boss-Broussard-Challenges-Facing-the-Preservation-of-Born-digital-News-Applications.pdf.

[14] Fernando Chirigati et al., "ReproZip: Computational Reproducibility With Ease" (ACM Press, 2016), 2085–88, doi:10.1145/2882903.2899401.

[15] Jeffrey R. Spies, "The Open Science Framework: Improving Science by Making It Open and Accessible" PhD diss., University of Virginia, 2013, https://osf.io/t23za/.