Abigail Goben and Dorothea Salo

Federal research

Data requirements set to change

FERPA, HIPAA, FOIA, and other sunshine laws, National Science Foundation data-management plans[1]—grant-funded research data has had compliance strings attached for some time. Attention to research data is now even more heightened following the responses of the federal agencies in August to the Obama Administration's Office for Science and Technology Policy (OSTP) directive from February 2013.[2] Research libraries will need to educate and partner with researchers to improve understanding and compliance, promote proper archiving of digital data, and expand discovery and reuse of research datasets.

History

Previous federal legislation governing data from funded research focused on maintaining privacy and security. Examples include the national security requirements surrounding data for Departments of Defense and Energy grants, as well as the stringent requirements facing federal research subcontractors under the Federal Information Security Management Act (FISMA).

Perhaps the most broadly known example of data-related security legislation is the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule of 1996. With mandated compliance after 2003, protected health information, including everything that could allow for personal identification of a patient from their data, was now regulated for use, reuse, and disclosure. When passed, this had significant impact on the accessibility of health data to researchers, with reports of greatly increased costs, time burdens, and difficulty in obtaining research data.[3] Research institutions currently seek new ways to obtain de-identified health information for greater researcher access, a process that may spur HIPAA reform.

The National Institutes of Health (NIH) for some time boasted the only major federal data-sharing mandate: "Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible."[4] This requirement has been honored more in the breach than the observance, clinical confidentiality often serving as an all-purpose reason not to share.

Beginning January 18, 2011, NSF required that all grant applicants submit a two-page research data management plan. While best practices for data management were already well established in some fields (e.g., earth science, psychology), many disciplines began to consider for the first time what those data management plans should include for them.[5]

Abigail Goben is assistant information services librarian at the University of Illinois-Chicago Health Sciences Library, e-mail: agoben@uic.edu, and Dorothea Salo is faculty associate at the School of Library and Information Studies at University of Wisconsin-Madison, e-mail: salo@wisc.edu

Contact Claire Stewart—series editor, head of digital collections and scholarly communication service at Northwestern University—with article ideas, e-mail: claire-stewart@northwestern.edu

While the mandate required the data plan, no specific requirements for best practices or gold standards were included, nor was data sharing mandated NSF-wide, though individual NSF divisions and directorates may mandate it, and some (e.g., Earth Sciences) do. The general expectation of the research community was that as best practices became apparent to grant reviewers, standards for the data management plans would increase, as would the impetus to share data.

The Data Management Plan Tool, from the California Digital Library and partners, is one tool that has been created to assist researchers in templating data management plans.[6] While no researchers have stepped forward to list their failure to obtain a grant due to a poor data management plan, anecdotal data suggests that reviewers have passed where the plan did not meet new and rising expectations.

Shortly after the NSF mandate came the highly publicized pushback from the research community against the Research Works Act (RWA) proposed to the 112th US Congress in December 2011.[7] This publisher-driven bill primarily focused on academic research articles published in peer-reviewed journals, and was poised to revoke the 2008 NIH Public Access Policy[8] as well as prohibiting other federal agencies and colleges and universities from requiring open access from their grantees and employees.

Researcher Heather Piwowar pointed out that the bill also included sweeping language that would have subsumed "all published" research datasets (including those in tables, supplementary information, and presumably nonfederal data archives)"[9] in the open-access prohibition. Ultimately, in response to the overwhelmingly negative reaction of the research and education community, the sponsors withdrew their support for the bill.

In direct response to RWA, the Federal Research Public Access Act (FRPAA) was re-introduced to Congress in February 2012. This act, however, focused specifically on the final outcome—the journal articles—produced from the funded research of 11 federal agencies. In specific regards to data, the bill states "laboratory notes, preliminary data analyses, notes of the author, phone logs, or other information used to produce final manuscripts" were to be excluded from the mandate.[10] Despite support, this bill was referred to committee, from which it did not emerge before the end of the congressional calendar.

OSTP memo

An even greater grassroots response than the opposition to RWA emerged from the May 2012 launch of a White House petition entitled "Require free access over the Internet to scientific journal articles arising from taxpayer-funded research." This was one of the first petitions to face the then-new requirement of reaching 25,000 signatures during the 30-day window, a milestone achieved in barely more than one week. While this petition primarily targeted access to the journal articles produced through scholarly research, the ultimate response to it also focused on data.[11]

The White House's February 2013 response to this petition[12] and comments from the research and library communities gathered by OSTP between November 2011 and January 2012[13] helped form the eventual policy memorandum from the White House and Obama Administration through the OSTP.[14]

This document, also released in February 2013, instructed the heads of executive departments and agencies with a research and development budget of more than $100 million annually to develop policies and plans to disseminate publicly funded research openly. The memo does not solely focus on journal articles, ending its first paragraph with the clear statement that "such results include peer-reviewed publications and digital data."[15]

Further, section 4 of the document outlines the specific objectives for both preserving research data and ensuring that it becomes speedily accessible, within boundaries of privacy concerns, national security, current law, etc. The memo gave agencies six

months to develop specific procedures and report them to OSTP; draft plans were due in August 2013.

A further White House Executive Order was then issued in May 2013,[16] which required that federal agencies "collect or create information in a way that supports downstream information processing and dissemination activities."[17] The requirements in the document included open formatting, usable metadata, data standards, and machine readability. The agencies were also charged with creating data inventories with the focus of providing a clearer picture of what data could be shared and improving government transparency.[18]

In June, many major publishers put forth a proposal called CHORUS to address the open-access requirements spelled out in the OSTP memo.[19] The Association for Research Libraries (ARL), the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU) issued their own proposal, SHARE, shortly thereafter.[20] Neither proposal fully addresses the data-sharing requirements outlined in the OSTP memo, and federal agencies are under no explicit onus to accede to, or even heed, either proposal.

Libraries and open data

Open sharing of data has previously varied widely by discipline. Data sharing tends to be more common with expensive-to-gather data such as astronomy, meteorology, or certain kinds of earth science data, while it is less common with medicine or library-science data.

One well-established example of required data sharing is GenBank from the National Center for Biotechnology Information (NCBI). While researchers are permitted to delay public access to submitted sequences in GenBank for a reasonable amount of time in order to publish their findings, the research and publishing communities expect sequences to be deposited promptly into GenBank, which currently holds more than 150 billion bases.[21] This expected sharing facilitated the speed at which the Human Genome project was first completed and has provided extensive medical benefits, such as the 2005 identification of an isolated case of polio in the United States.[22]

As further information about the drafts from the federal agencies coalesces, one clear theme emerges: increasing requirements for preserving and sharing federally funded research data and an associated increase in reuse of existing data. No matter whether federal agencies choose a solution resembling CHORUS, SHARE, or NIH's existing PubMed Central, these new challenges offer a number of opportunities for libraries and librarians.

As with the NIH Public Access Policy, alerting researchers and keeping campus administrators well-informed will be the first order of business. Liaison librarians are, as always, the natural conduits to faculty, while associate university librarians and university librarians will need to undertake communication with campus IT and high-level university administrators in collaboration with research offices. In disciplines where data sharing is not the norm, this communication is liable to be ticklish and difficult, as dismayed researchers worry about scooping, data licensing, expense, and the often-considerable added effort involved in readying data for sharing.

Should federal agencies converge on a solution resembling SHARE, institutional repositories and their managers will find themselves rudely thrust back into the limelight. Implementing SHARE would demand substantial immediate investment in technological improvements to institutional repository software. Repositories running on a skeletal staff complement, as many do, will also need considerable extra staff reinforcement, at least temporarily, if they are to withstand the sudden onslaught of faculty and external service demands. And due to the expected multiplicity of policies, liaison librarians will need continuing education in order to effectively collaborate with faculty to find the best data management and curation practices, understand

requirements for compliance, improve data discovery and reuse, and document data reuse and impact.

Whichever model federal agencies choose, the ensuing rush of open data will create considerable demand for data-specific reference and instruction, well beyond current emphasis on data-management planning. From data citation to data preservation to alternative metrics that take data production and reuse into account alongside commonly accepted (though flawed) measures such as journal impact factor, researchers at all stages in their careers will legitimately find themselves in need of exactly the kind of guidance academic librarians can offer.

Notes

1. "Dissemination and Sharing of Research Results," Division of Institution and Award Support, National Science Foundation, November 20, 2010, accessed June 25, 2013, www.nsf.gov/bfa/dias/policy/dmp.jsp.

2. Michael Stebbins, "Expanding Public Access to the Results of Federally Funded Research," Office of Science and Technology Blog, February 22, 2013, accessed June 25, 2013, www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research.

3. Institute of Medicine (US) Committee on Health Research and the Privacy of Health Information: The HIPAA Privacy Rule; S. J. Nass, L. A. Levit, L.O. Gostin, editors, "Effect of the HIPAA Privacy Rule on Health Research," *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* (Washington D.C.: National Academies Press, 2009), accessed June 25, 2013, www.ncbi.nlm.nih.gov/books/NBK9584/.

4. "Final NIH Statement on Sharing Research Data," February 26, 2003, accessed June 25, 2013, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

5. "Dissemination and Sharing of Research Results."

6. California Digital Library, DMPTool. 2010–2013, accessed June 25, 2013, https://dmp.cdlib.org/.

7. Darrell Issa and Carolyn Maloney, "H.R. 3699," Government Printing Office, December 16, 2011, accessed June 25, 2013, http://www.gpo.gov/fdsys/pkg/BILLS-112hr3699ih/pdf/BILLS-112hr3699ih.pdf.

8. National Institutes of Health Public Access, accessed June 25, 2013, http://publicaccess.nih.gov/.

9. Heather Piwowar, "Research Works Act Attacks Data Dissemination Too," *Research Remix,* January 7, 2012, accessed June 25, 2013, http://researchremix.wordpress.com/2012/01/07/rwa-data/.

10. Michael Doyle, Kevin Yoder, and William Lacy Clay, "H.R. 4004," Government Printing Office, February 9, 2012, accessed June 25, 2013, www.gpo.gov/fdsys/pkg/BILLS-112hr4004ih/pdf/BILLS-112hr4004ih.pdf.

11. "Require Free Access Over the Internet to Scientific Journal Articles Arising from Taxpayer-Funded Research," *We the People,* May 20, 2012, accessed June 25, 2013, https://petitions.whitehouse.gov/petition/require-free-access-over-internet-scientific-journal-articles-arising-taxpayer-funded-research/wDX82FLQ.

12. John Holdren, "Increasing Public Access to the Results of Scientific Research," *We the People,* February 2013, accessed June 25, 2013, https://petitions.whitehouse.gov/petition/require-free-access-over-internet-scientific-journal-articles-arising-taxpayer-funded-research/wDX82FLQ.

13. Office of Science and Technology Policy, on behalf of the National Science and Technology Council, "Request for Information: Public Access to Digital Data Resulting from Federally Funded Scientific Research," *Federal Register,* November 04, 2011, accessed June 25, 2013, https://www.federalregister.gov/articles/2011/11/04/2011-28621/request-for-information-public-access-to-digital-data-resulting-from-federally-funded-scientific.

14. Stebbins, "Expanding Public Access."

15. John Holdren, Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Fed-

erally Funded Scientific Research," February 22, 2013, accessed June 25, 2013, www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

16. Barack Obama, "Executive Order—Making Open and Machine Readable the New Default for Government Information," May 9, 2013, accessed June 25, 2013, www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

17. Sylvia Burwell, Steven VanRoekel, Todd Park, and Dominic Mancini, "Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy —Managing Information as an Asset," May 9, 2013, accessed June 25, 2013, www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

18. Obama, "Making Open and Machine Readable."

19. Andi Sporkin, "Understanding CHORUS," Association of American Publishers, June 5, 2013, accessed June 25, 2013, http://publishers.org/press/107/.

20. Prue Adler, Judy Ruttenberg, and Julia Blixrud, "Shared Access Research Ecosystem (SHARE) proposed by ARL, AAU, APLU," *ARL News,* June 7, 2013, accessed June 25, 2013, www.arl.org/news/arl-news/2773-shared-access-research-ecosystem-proposed-by-aau-aplu-arl.

21. Ilene Mizrachi, "GenBank: The Nucleotide Sequence Database: History," *The NCBI Handbook,* McEntyre J, Ostell J, ed. created: October 9, 2002; last update: August 22, 2007, accessed June 25, 2013. www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.History.

22. Kathy Cravedi, "GenBank Celebrates 25 Years of Service with Two-Day Conference; Leading Scientists Will Discuss the DNA Database at April 7-8 Meeting," *NIH News,* April 3, 2008, accessed June 25, 2013, www.nih.gov/news/health/apr2008/nlm-03.htm.

("LibQUAL coding cohorts," cont. from page 420)

Others liked the aspect of getting to know library employees from different parts of the library and are now engaged in networking and talking about library issues with them.

One coder indicated that it was "really energizing" and all coders enjoyed seeing the positive comments as well. Two coders who said they would not participate again felt that trained students could do the coding or said they would only do it if they felt obligated. One mentioned, "I'll do it to be a team player, but I would rather not." Two were neutral about future participation.

Conclusion

The focus group comments helped us to see where we could improve the process the next time around. Conversations with the library's Administrative Council helped us determine the methods of sharing the information library-wide. Plans were made to disseminate the findings as widely as possible throughout the library and provide discussion venues for all employees. The comments in their entirety were posted on the library's internal wiki, along with comments arranged by the discipline of the survey participant.

Additionally, a session during the fall library retreat was devoted to working with the comments in groups. The group discussion reaped even more ideas from library personnel in regards to addressing the concerns of library patrons. Even though we have done a lot of work with the comments, the process is not complete. We will hold a discussion in a library-wide meeting to prioritize problems needing solutions. Thus far the project has served as a unique learning experience for the library and has proved to be of great benefit to all library employees involved.

Note

1. Daniel O'Mahony, "Sharing the Wealth: A Process for Engaging a Large Group in Coding LibQUAL+® Survey Comments," poster presented at the 2010 Library Assessment Conference, Baltimore, Maryland, October 24–27, 2010.