2019 IFLA WLIC Athens

Submitted on: 02.07.2019

Public Access: A Driver for Preservation and Discovery of Datasets at a US Land-Grant Institution

Andrea L. Ogier
University Libraries, Virginia Tech,
Blacksburg, VA USA
alop@vt.edu

Jonathan Petters, PhD
University Libraries, Virginia Tech,
Blacksburg, VA USA

Virginia Pannabecker
University Libraries, Virginia Tech
Blacksburg, VA USA

Robert Settledge, PhD
Advanced Research Computing, Virginia Tech
Blacksburg, VA

Elizabeth Grant, PhD
School of Architecture + Design, Virginia Tech
Blacksburg, VA

Samantha M. Harden, PhD
Department of Human Nutrition, Foods, and Exercise, Virginia Tech
Blacksburg, VA

Julie Griffin
University Libraries, Virginia Tech
Blacksburg, VA

Tyler O. Walters, PhD
University Libraries, Virginia Tech
Blacksburg, VA

Abstract:

Public access to federally funded research data ensures preservation and discovery of datasets to promote translation of research evidence into meaningful outcomes. However, historical policy and concerns regarding making data publicly accessible impede realization of implementing public access to data. These concerns include worry over the treatment of intellectual property, the cost (in time and money) of making research publicly accessible, and the danger of accidentally releasing human subjects data. To overcome these issues, a Public Access to Data Committee was established at a public university in rural southwest Virginia. In this paper we review the history of federal public access provisions, share goals, and describe the committee's process to ultimately engage faculty and administrators in this critical link from research to practice.

Keywords: Public Data, University Governance, Federal Mandates.

Introduction

Virginia Tech takes seriously its history as a land-grant public university, emphasizing its role in creating innovative and positive change in the local community, throughout the Commonwealth of Virginia, and across the country and globe. The University's living motto, *Ut Prosim* (That I May Serve) embodies this spirit of service to the world. However, Virginia Tech is also an R1 research institution that supports thousands of researchers spread throughout seven research institutes and nine colleges, and has half a billion US dollars in private contracts and federal grants. While providing public access to research data is vitally important to its land-grant mission of using research for the public good it is also a controversial and costly subject for researchers. To support researchers who need to comply with federal mandates on public data we established a Public Access to Data Committee charged with exploring the current landscape of public data and making recommendations to improve support of public data at Virginia Tech. This faculty-led committee is sponsored by the university's Commission on Research, and includes representation from administrative, teaching, and research units across campus. Our approach is driven by the idea that preserving and providing access to data directly supports our university's stated goals of bridging the divide between research and practice. Further, we emphasize the researcher-centric view that the university has developed (and should continue to develop) services that remove the burden from the researcher. This paper reports on our ongoing efforts coordinating and leading this committee and delivers the results of our semester-long exploration of university mechanisms that support public access to data.

A Short History of Public Data

Public access to data (or public data) can be a contentious subject, even more so than "open data" and "Findable, Accessible, Interoperable and Reusable (FAIR) data". While the terms "public data" and "open data" are sometimes used interchangeably, in this case we use the term "public data" and "public access to data" to refer to data funded by a federal government, either within its agencies or via grants to researchers, that is made accessible to the public. The history of public research data in the US broadly begins in 1999 with a rider, called the Shelby amendment, that was attached to Public Law 105-277. Meant to improve transparency and accountability of data underlying EPA regulations, the Shelby amendment instructed the Director of the Office of Management and Budget to "require Federal awarding agencies to ensure that all data produced under an award will be made available to the public through the procedures established under the Freedom of Information Act." This rather broad mandate had significant repercussions, as scientists and researchers--who depend on federal grants--came down on both sides of the issue (Frankel 1999; Fischer 2013). According to Eric Fisher, Senior Specialist in Science and Technology of the Congressional Research Service, those in favor of the amendment argued that the public, not just government scientists and peer reviewers, should have access to data underlying regulatory measures because 1) the data could be incorrect, 2) regulations could have a broad area of effect on the public at large, and 3) implementation of these measures could be extremely costly to the public (Fisher, 2013). Opponents cited concerns over human subjects data, the undue burden

placed on the research enterprise, and the cost of making data publicly accessible. In a 1999 opinion piece contributed to *Science*, Mark Frankel questions the effects of the Shelby amendment on scientific progress, and specifically calls attention to intellectual property:

> How will intellectual property rights be accommodated by the new requirement? Under U.S. law, scientists have a year from the date of publication to file a patent application. Will allowing data to be publicly available through [the] FOIA threaten a scientist's foreign patent rights? How will the revision affect university-industry partnerships, if such collaborations involve a commingling of private and public monies? Will ambiguities in determining which data would be subjected to a FOIA request make industry reluctant to pursue such collaborations?

Much of the fear over the Shelby amendment involves the possibility of rival researchers making FOIA requests--*Nature biotechnology* even goes so far as to publish a short article aimed at helping researchers shield their innovations from FOIA requests (Eisenstein and Resnick, 1999). These discussions around the Shelby amendment show the central tension in public data: how to create accountability, create transparency, and foster innovation without risking exposure of human subjects, allowing competing researchers access to novel ideas (i.e. research "scooping"), and stifling scientific progress. Twenty years after the Shelby amendment, the public access to data discussions started by the AAU (Association of American Universities) and APLU (Association of Public and Land-grant Universities) still center on how to balance these two perspectives.

While the National Institutes of Health implemented a data sharing policy in 2003 and the National Science Foundation has required data management plans since 2011, the next big step forward for large-scale policy around public access to data is known colloquially as the "2013 OSTP Memo" which was drafted by John Holdren, Director of the Office of Science and Technology Policy within the Obama White House. This memo and subsequent efforts were legislated by Congress in the America COMPETES Reauthorization Act of 2010 (H.R. 5116). While the concept of making "digital data" and peer reviewed publications available to the public existed before 2013, the OSTP memo explicitly directed each federal agency with over $100 million in research expenditures to develop (though not necessarily enact) plans to support greater public access to the results (including data) of federally funded research. The release of this memo led to the creation of both article and data sharing policies for most federal research agencies.[1]

However, policies, requirements, and measurable compliance with those requirements are three different things. While these federal agencies have largely disseminated policies that require grantees to consider public access to their data, implementation of compliance measures and metrics for measuring the value of public data have not yet been widely established. Nor has the cost of making data publicly accessible been adequately addressed, though agencies do allow costs associated with data management to be charged to the grant. However, since the indirect rate of many grants to institutions of higher education is at or over 50%, including these costs in the direct portion of the grant could significantly hamper the research process. Making data useful and publicly accessible requires significant resources in both time and money, neither of which are readily available to researchers who are balancing teaching, research, and tenure requirements.

Similarly, neither the agencies nor the federal government itself specifically address the technical and social infrastructure needed to support public access to data. In a 2013 policy piece in *Science*, Francine Berman and Vint Cerf discuss many challenges surrounding implementation of the OSTP Memo's requirements, including the monetary cost of building infrastructure, the skills and processes needed to steward diverse datasets over time, and the difficulty in transitioning between private and public sector

[1] See SPARC's catalog of data sharing policies at http://datasharing.sparcopen.org/compare?ids=9&compare=data

infrastructure support. At the end of the discussion, Berman and Cerf suggest that "the key is not to look to a particular sector alone but to develop much stronger partnerships among sectors." Furthermore they recommend incentivizing private companies to take an interest in stewarding research data (with appropriate checks on access and use), clarifying public-sector interests and resources for supporting public access, and supporting a "research culture change" to encourage researchers to take advantage of private sector resources. Ultimately, Berman and Cerf draw attention to the importance of practical economic models to sustain the infrastructure needed to make federally funded research data public and available over the long term.

In 2017 the American Association of Universities and the Association of Public Land Grant Universities convened a working group to explore this space and make recommendations to both federal agencies and institutions of higher education. The Public Access Working Group report begins by asserting that public access to things, be it research findings and/or data, serves the public good.

> In this era of open scholarship, greater access to research findings and data [...] has proven to be an important way to accelerate scientific progress and advance innovation to better serve the public good.

Furthermore, it advises that

> Universities will need to create the infrastructure required by the public access mandates of the federal agencies funding their research so that data collected to support federally funded research can be shared, to the extent possible, with the public.

Shifting the cost of creating infrastructure to support public data to the university could offset much of the burden on researchers, allowing them to more efficiently comply with federal policies. However because every university has a different organizational structure and research ecosystem, neither of these reports present practical strategies for helping individual researchers walk the delicate line of compliance with federal mandates requiring transparency and accessibility while still protecting human subjects, secure datasets, and intellectual property. Recognizing this, the AAU and APLU partnered with the NSF to hold a two-day Public Access to Data Workshop in October 2018.

Three representatives from Virginia Tech (from the Library, Commission on Research, and Provost's Office) attended the AAU-APLU's Public Access to Data Workshop meeting in Washington DC. The workshop had three goals: 1) increase the adoption of policies and infrastructures that support public access to data at research institutions, 2) encourage inter-institutional collaboration to support this effort, and 3) provide a space for institutional representatives to draft, discuss, and compare action plans. As a deliverable from this workshop, the Virginia Tech team drafted two action plans, one aimed at gaining institutional support for public data through governance, the other to train researchers (both graduate students and faculty) on relevant policies, processes, and infrastructures (both local and national/international) that lower the barriers to making data publicly accessible. While both plans could be enacted concurrently, the team decided to work through the governance system and identify gaps and pain points that should be addressed in the creation of resources and services that lower the barriers for researchers at VT. At the time of publication, the governance action plan is wrapping up, and the rest of this article presents the findings that will be used to inform the training action plan.

Creating the Public Access to Data Committee at Virginia Tech

Following the action plan developed at the workshop, the VT representatives began drafting a charge to the Commission on Research for a representative committee to explore existing resources that support researchers who need to provide public access to their data, and to identify gaps that could be filled by future resources, services, or infrastructures. The charge sets out three exploratory areas which are then tied together into a final report with recommendations. The three areas of exploration are: 1) services

and support available to researchers in sharing their data, 2) a review of relevant policies at Virginia Tech that govern researchers' ability to share their data, and 3) a review of Data Management Plans (DMPs) from active research studies tracked by Office of Sponsored Programs. The final report with recommendations will be presented to the Commission in Fall 2019.

Virginia Tech's governance structure is organized into commissions and committees, each of which are authorized to charge short term working groups to make recommendations and provide input on relevant topics.[2] While working through governance can take longer than convening an ad-hoc group of like-minded individuals, it has the benefit of lending the authority of governance toward gaining input from representatives across campus whose perspectives on public access to data may not be immediately apparent. The charge for this group was predicated on assembling a diverse group of people each with a unique viewpoint. Approaching this topic through governance allowed us to seek input from a variety of sources across campus. In addition, it offers a potential path forward for the creation of a university-level resolution, should a new policy or a change in policy be deemed necessary. The charge was sent to high-level administrators for feedback in November, presented to the Commission on Research, and approved on December 12, 2018.

Approval from a university-level commission in hand, membership for the Public Access to Data Committee at Virginia Tech (hereafter termed 'the Committee') was quickly established with representatives from across the university, including those from the Division of Scholarly Integrity & Research Compliance, the Office of Sponsored Programs, the Division of Information Technology, Advanced Research Computing, University Libraries, the Commission on Research, Colleges, research Institutes, the Faculty at large (including tenure-track, research faculty, and collegiate faculty), the Graduate School, and University Legal Council. Monthly meetings were established and each focused on a different deliverable.

Discussion Summary

Deliverable 1: Data sharing services and support at Virginia Tech

Research services and support can come from a variety of sources: the institutes themselves, colleges or departments, the Office of Research and Innovation, the University Libraries, and IT/Advanced Research Computing. Each of the units named above has their own areas of concern and expertise: many have developed resources for specific subsets of the research population. Bringing together a group of representatives from each unit on campus allowed us to discuss what services and support exist, and what would be helpful to create.

Clearly articulated needs included:

- Custom image sharing services (i.e. a way to create a browseable and searchable custom digital library to share images either publicly or with specific collaborators)
- De-identification support for human subjects data or data containing personally identifiable information (PII)
- Making data more discoverable for a particular research community
- Guidance around what types of data can and cannot be made publicly accessible or shared, including data at different steps in a research workflow (raw data vs analyzed data vs aggregate data)
- Methods for sharing data that are not repository-based (i.e. emailing data to another collaborator)
- Assistance with creating data use agreements for sharing datasets that should be shared but cannot be made publicly accessible
- Guidance on how and when to use a secure research environment to work with and provide access to healthcare data and electronic medical records

[2] See Article XI: Other Committees on https://governance.vt.edu/bylaws---constitution.html

Committee discussions centered on the kinds of data that can and cannot be made public. Health care and medical researchers, for example, cannot make public any data that can be linked back to a specific person or easily identifiable group without their consent. However, there may be aggregate data that (if properly de-identified) could be shared or made public. Institutional Review Boards (IRBs) can provide some guidance for human subjects data, though there was some question about whether human subjects data that is IRB-exempt could be made public. The cost of providing public access to data was also a point of discussion, though given the prevalence of both institutional repositories and "free" repositories like figshare, the concerning cost was in labor and time rather than in monetary resources.

One identified challenge is the diverse nature of the datasets, another is that the landscape of policy and funding agency mandates is continually changing. The Common Rule, for instance, now requires that IRB approved consent forms be posted publicly (Common Rule §46.116 subsection h). How should these forms be posted and for how long? Should they be preserved alongside the peer-reviewed article and the dataset in an institutional repository? The committee agreed that it is too much to ask researchers to stay up-to-date on these policies and practices; instead, this is an area where the research institution should step in and provide current, just-in-time guidance and support.

While the committee was able to identify a number of research-related, researcher-centered services provided by the University, only the repository services and guidance provided by the Library actively supported public access to research and research data. Services and guidance provided by both Research Compliance and the security office within the Division of IT are concerned with keeping data safe and secure, while the Library's guidance covers how and why data should be made available. In the end, the committee recommends that the Library, IT, and Research Compliance work together on creating guidance to help researchers know when data should or could be shared, and when it must remain secure.

Deliverable 2: Relevant policies in effect at Virginia Tech

For the second deliverable, the committee reviewed relevant policies governing research data at Virginia Tech to determine whether they provided enough guidance to help researchers know when and how to share their data. The committee was able to identify six official VT policies[3] that have some bearing on this topic (13000, 13015, 7010, 1060, 7100, and 7000), though only 13000 and 13015 address intellectual property and research data. Frankel's observations about FOIA requests, intellectual property, and the Shelby Amendment remain a significant part of this discussion, as do the public access provisions required by the OSTP Memo.

The Policy on Intellectual Property (Policy 13000, revised 2015) is designed to establish "ownership criteria" and resolve questions surrounding ownership of intellectual property, and applies to "all employees, students, and all other persons or entities using University resources." Although this policy does not clearly consider research data to be a form of Intellectual Property, it does explicitly cover all other research products (research papers, books, software, inventions, articles, etc.). While this policy does explicitly state that "many IPs are best disseminated by publication and placed in the public domain" it also asserts that "a significant number" should instead be protected by IP law "with attendant financial considerations." The policy further distinguishes between "traditional results" which the author owns, and "novel results" of research which are created as a condition of employment, and which hold some significant benefit (i.e. monetization or technology transfer) to the University. In the latter case, the University asserts ownership. Interestingly, sections 2.3.A.3-4 cover sponsor rights, which are interpreted as private sector sponsors (not federal agencies) and federal agency rights, which are interpreted as statutory IP rights to patents. Neither of these specific cases, nor any other in the document, clearly address the federal requirements for public access to research results which may be considered to be IP by the University and fall under the purview of the Shelby Amendment and OSTP Memo. The question of whether research or data funded by a federal agency, generated by a private

[3] See VT Policy Library at https://policies.vt.edu/policy-library.html

university, and considered to be significant intellectual property is subject to a FOIA request remains unanswered. Policy 13000 states that the federal agencies have statutory IP rights to any patents generated, but does not clearly provide guidance for research data or related materials.

On the other hand, Policy 13015, Ownership and Control of Research Results (revised 2015), does explicitly apply to research data and asserts that the university has ownership of "research results and material (this includes all data)" generated with university resources. However, this policy does not differentiate between research materials that are wholly funded by the university (through researcher salary and facilities space) and those that are partially funded by federal agencies through direct grants and indirect (facility and administration) rates. Although data funded by federal money granted to the University are "owned" by the university, per this policy, there is no guidance as to whether those data are also subject to federal access policies and provisions.

Deliverable 3: Review of awarded Data Management Plans (DMPs) tracked by the Office of Sponsored Programs (OSP)

As public access to data is generally tied closely to federally funded research, reviewing data management plans from funded applications written by VT researchers helped the committee get a better sense of existing gaps in researcher and institutional knowledge around data management best practices. More thorough studies of data management plans submitted to NSF have been performed at other universities (see Mischo, et al; Parham, et al), but for the purposes of this committee, a subset of DMPs from recently awarded grants by a variety of federal agencies were anonymized and made available to the committee for review. Rather than considering whether the data management plans were "good" or "bad," the committee instead discussed whether the author had been able to find appropriate guidance about local or disciplinary services and whether the university and researcher would be able to enact the data management plan as written. From a legal standpoint, federal grants are awarded to Virginia Tech, not the individual PI, thus, any change to the documents submitted in the application must be approved through the university. A funded application is considered a legally binding document, including the data management plan. Thus any modification or change to the plan, including the DMP, needs to be approved through OSP.

The award lifespan for a federal grant can be anywhere from 1 to 5 years, even without adding on the additional 1 or 2 years between submitting the application and awarding the grant. In the intervening 2 to 7 years technology, services, support, and infrastructure can change rapidly and without warning. In the reviewed DMPs, the committee found a number of dead websites referenced as preservation and access mechanisms, including one or two that had, at one time or another, been maintained by the university. Similarly, outdated technologies were referenced, and quite a few researchers were either not aware of the federal requirements for public access to federally funded data, or considered making the data available on a case-by-case basis "upon request" to be sufficient. While there certainly are instances where "upon request" is warranted, the burden of making such a justification in the DMP falls on the researcher. The DMPs the committee investigated did not provide this justification. Often researchers start from a position of secure or private data, many times for good reason, either because of privacy concerns, requirements from IRB, or because of NDAs (non-disclosure agreements) from the private sector. However, often data that do not contain IP, PII, or restricted assets are not made publicly available because of a fear of "scooping," because of changes in federal mandates, or simply because researchers are not informed about the difference between data that can be made public and data that cannot be made public.

Over half of the reviewed DMPs could be fully enacted either as written or with minor revisions given current services and technologies in existence at Virginia Tech and elsewhere. The remainder, however, are in need of clarification or major revision. Beyond the comments discussed above, common issues include confusion over whether data can be disseminated via traditional journal and conference publication methods (in most cases it cannot), lack of clarity around data types (raw vs analyzed vs publication-ready), and lack of knowledge of standard data publication and preservation practices (i.e.

stating that the High Performance Computing servers have long-term preservation and access mechanisms, which they do not). Although the committee agreed that these issues are problematic from a compliance standpoint, they also agreed that researchers should not bear the burden of these revisions on their own.

Recommendations: Framing University Support for Public Access to Data

The Committee's discussions, when viewed as a whole, reveal a few gaps in Virginia Tech's institutional support for public access to data. It is important to note, however, that the burden to address these gaps should not fall on one single department or administrative unit. Instead, the Committee strongly recommends that the university take a holistic approach to supporting public access to data. As Berman and Cerf note above in spurring this support:

> The key is not to look to a particular sector alone but to develop much stronger partnerships among sectors. Such a division of labor can provide a framework of options that distribute the burdens and benefits of stewardship and economic support. (Berman and Cert)

Stronger partnerships between units engaged in research support would lead to faster and more efficient solutions for common problems, would ensure that researchers have the information they need when they need it, and would lead to stronger grant proposals. The cost of making data publicly accessible is significant; if enacted strategically, this burden could be divided across the researcher, sponsored programs, research compliance, IT, and the library. However, creating these partnerships takes time and resources, both of which are usually in short supply. How to incentivize, assess, and value these partnerships at a high level is still an open question for the Committee (and possibly beyond its purview), but we recommend that the Commission on Research consider ways to better encourage the formation of these partnerships. We will also seek ways to recommend to national representative groups, such as the AAU and APLU, and to national funding agencies, such as the National Science Foundation, that a key way to incentivize implementation of policies to make data public would be to establish funding programs specifically for universities to develop infrastructure and services that support their researchers in making their data publicly available. As the greatest cost burden often in initial establishment of such programs, such funding mechanisms could help universities operationalize their goals in making research data public.

Both of the above recommendations could also apply to research support across the university. It was interesting for the Committee to note that several members were involved in a parallel discussion on restricted and secure data. While public access and restricted access are normally opposing endeavors, we found that researchers had the same types of questions about restricting data as they did about making data public. Thus, another, more specific, recommendation is that the Library, IT, and Research Compliance work together on creating a guidance system to help researchers know when data should or could be shared, and when it must remain secure.

Supporting researchers who want or need to make their data publicly accessible inarguably contributes to Virginia Tech's land-grant mission. Publicly sharing data and other research products can lead to improvements in the public good. However, knowing when, what, and how to make things publicly accessible is much less clear. Researchers need their institutions to provide this guidance and these services. We know that Virginia Tech will rise to meet these challenges in the spirit of its living motto, *Ut Prosim* (That I may serve).

References

AAU-APLU Public Access Working Group. (29 Nov 2017). Report and Recommendations. Retrieved from https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf

Berman, F., & Cerf, V. (09 Aug 2013). Who Will Pay for Public Access to Research Data? *Science*, vol. 341, issue 6146, pp. 616-617 doi: 10.1126/science.1241625

Eisenstein, R.I. & Resnick, D.S. (Jun 1999). Public access to data may be closer than you think. *Nature biotechnology*, 17(6):604-605.

Fischer, E. A. (2013). Public Access to Data from Federally Funded Research: Provisions in OMB Circular A-110 Note. *CRS Report for Congress*. Retrieved from https://fas.org/sgp/crs/secrecy/R42983.pdf

Frankel, M. S. (19 Feb 1999). Public Access to Data. *Science*, 19 Feb 1999: Vol. 283, Issue 5405, pp. 1114 DOI: 10.1126/science.283.5405.1114

Holdren, J. P. (22 Feb 2013). Increasing Access to the Results of Federally Funded Scientific Research. Retrieved from https://www.science.gov/docs/ostp_public_access_memo_2013.pdf

H.R. 5116 -- 111th Congress. (17 Dec 2010). Retrieved from https://www.congress.gov/bill/111th-congress/house-bill/5116

Mischo, W. H., Schlembach, M.C., & O'Donnell, M.N. (2014). An Analysis of Data Management Plans in University of Illinois National Science Foundation Grant Proposals. *Journal of eScience Librarianship* 3(1): e1060. http://dx.doi.org/10.7191/jeslib.2014.1060

Parham, S.W., & Doty, C. (2012). NSF DMP Content Analysis: What Are Researchers Saying? *Bulletin of the American Society for Information Science and Technology*, 39, no. 1: 37-38. doi:/10.1002/bult.2012.1720390113

We Paid for the Research, So Let's See It. [Editorial]. *New York Times*, 25 Feb 2013. p.A24. Retrieved from **www.nytimes.com/2013/02/26/opinion/we-paid-for-the-scientific-research-so-lets-see-it.html?_r=0**.

PUBLIC LAW 105–277 (21 Oct 1998). Retrieved from https://www.congress.gov/105/plaws/publ277/PLAW-105publ277.pdf p496

White House. CIRCULAR A-110. (amended 30 Sept 1999). Retrieved from https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-110.pdf