Advancing research through data management

By Trevor Riley

New funding agency rules for open-access data drive the need for data literacy and data management skills.

Research data management issues are not simple. They are not solved solely through actions of the researcher. Solutions require involvement of many stakeholders within government, industry, academia, and international organizations. The current top-down approach by government agencies serves only to address a narrow band of research performed. To make real progress, we must take additional steps to address cultural barriers, implement best practices, and instill the importance of data management and the value of data sharing in the next generation of researchers.

Change to come

In the past decade, the White House and federal agencies have faced growing expectations for greater transparency and accountability. Although this pressure is not new, technology has developed to the point where citizens can lobby for a cause they believe in by signing an online petition. One such petition, which was posted in May 2012, targeted the openness of scientific journal articles resulting from taxpayer-funded research.[1] This petition, along with sentiment surrounding the nature of scholarly publishing, pushed the White House toward a monumental first step that will have major impact on the openness of scientific research.

In 2013, nine months after the petition was posted, the White House Office of Science and Technology Policy (OSTP) released

Capsule summary

DATA MANAGEMENT RULES CHANGE

In the United States, a citizen-driven petition in 2012 led to a federal mandate for open access to taxpayer-funded research, including research data. Agencies that award more than $100 million in research funding per year require principal investigators to provide data management plans to meet the requirement.

A COMPLEX MANDATE

Each funding agency wrote its own requirements for satisfying the data access requirement, although some agencies worked together to develop consistent requirements. However, data storage and retrieval infrastructure may be insufficient at some institutions. Finally, the research community currently lacks effective organization and management of large data sets.

DEVELOPING DATA MANAGEMENT SKILLS

Fields such as astronomy have developed tools for archiving and working with large data sets. Librarians, too, have developed programs for data information literacy skills training programs. Application of these tools and skills can help meet data management requirements.

a memorandum to the heads of executive departments and agencies.[2] The memo directed funding agencies with annual research expenditures more than $100 million to develop plans specifyimg objectives for publications and data to increase public access to research outputs related to federal funds. The memo allowed agencies room to tailor individual plans and required each agency to ensure that researchers submit data management plans along with proposals. Although some agencies, including the National Science Foundation, already met this requirement because of prior policies, the memo also included language that required agencies to develop a system to evaluate data management plans for merit. Further, agencies were asked to plan strategies for measuring and enforcing compliance. As of October 2015, 15 of the 21 agencies covered by this memo submitted plans, each at various stages of implementation.

Credit: Scholes Library, Alfred University

Scholes Library at Alfred University College of Ceramics. Librarians have a long history of organizing data in many formats. Today, librarians lead the effort to teach students and researchers about effective organization and management of data to meet federal data management mandates.

National Science Foundation

In March 2015, NSF published its public access plan: "Today's data, tomorrow's discoveries."[3] The plan responds to objectives raised in the OSTP memo and expands on previous policy, while pointing out that it plans to implement further requirements in future stages. Effective January 2016, all new awards granted by NSF are required to make publications, such as peer-reviewed articles, freely available no longer than 12 months after publication. Investigators are responsible for submitting publications to an NSF-designated repository and noting these publications in annual reports along with the Digital Object Identifier (DOI) for linking.

Currently, NSF has identified only the Department of Energy's Public Access Gateway for Energy and Science (PAGES) portal for submissions, but NSF will add to this list or work with the DOE to provide linking through original Versions of Record (VOR). Although this new requirement seems obvious to expand research output and should be easy for researchers to meet, it must not be overlooked. Federally funded research now is made freely accessible to all, and it no longer will be hidden behind publisher paywalls.

NSF is less straightforward regarding the requirements for other research outputs, including data. The language of the plan seems to postpone the question of data until a later time, while referring back to its original requirements on data management plans. There are many possible reasons why data is not as strongly addressed, but it is clear that the main challenge is development of an underlying infrastructure for long-term preservation and access through a network of institutions, publishers, and government agencies. Other reasons that likely have contributed to a staged approach, relative to data, relate to the inability of principal investigators to comply with more rigorous data requirements at this time. This inability results from lack of understanding of requirements, lack of skillsets to conduct proper management, and other limitations related to technological and support services.

Department of Energy

The Department of Energy published its public access plan in July 2014 and implemented it agency-wide in October 2015. As required by the OSTP memo, all proposals must contain a data management plan. DOE PAGES acts as a full-text database for submission of manuscripts, but researchers also have the ability, instead, to submit the link

of the publisher's VOR along with metadata and DOI. The department's access plan goes beyond the NSF plan and specifically states that data management plans will be evaluated on their merits and that failure to comply with what has been written in the proposal's plan will negatively influence future funding opportunities.

Language in the DOE plan shows that the department strongly believes in data management planning as a key part of the research process. The DOE plan is unique in its approach to education: It recognizes that skills related to effective management of data are tied to training and education. The department reaffirms its support for programs aligned with data management training for researchers at all levels, undergraduate to postdoctoral.

Turning toward education

There is not a single (accredited) engineering or materials program in the United States where a student can earn a degree without taking a course in calculus. This also is the case with other fundamental courses, such as chemistry, physics, and basic oral and written communication. Although we have invested in technical infrastructure through development of digital repositories, we have given less attention to integration of good data management practices into current research and curriculum. Throughout the sciences, a large majority of students graduate without basic understanding of data management or the ability to apply best practices in their research. We must remember that, although a robust infrastructure is important, without a clear understanding of how best to document, name, format, store, and share data, even the most capable researchers will continue to encounter the same issues of unreliable, untrustworthy, unintelligible, and unusable data.

Although research data are central to the communication of ideas so that others may test and verify, or disprove, results, data management has become an appendix—an addition to research. Exchange of knowledge solely through the traditional publishing model has become inefficient in data-intensive fields and no longer meets the needs of researchers. Although there is great value in production of scientific literature, the benefits of accessible, well-documented, and trustworthy data is more vital than ever.

Impact in academia

The lack of basic data management skills at the undergraduate level extends into graduate programs and industry, where an understanding of data management practices is arguably more important. We generally understand that graduate students in the sciences are relied upon to process, gather, and interpret data. We also accept that much of the research is in direct relation to or in support of faculty research. A recent study that examined the perspectives of graduate students in relation to data management found that practices were largely tied to previous experiences or training. The study also found that students had an overall lack of understanding of documentation and organization of data. The study authors bring this issue into focus by connecting these two points, saying, "given their close proximity to the data, the perceptions and attitudes of graduate students toward data management issues and the actions they take (or do not take) throughout the data lifecycle are likely to have a sizable impact."[4]

As a whole, academia is struggling with the implementation of data management. One of the most complete studies on research data management to date goes as far as to say that "virtually no one in academia perceives that they have a professional responsibility or mandate for research data management functions" and cites a lack of professional training as a major deficiency in effective management of research data over the long term.[5] Another look at major research institutions with significant grant activity found that faculty are generally unfamiliar with even their own institution's data management requirements.[6]

At the graduate level, concern is that behaviors and perceptions of faculty, who do not follow best practices or engage in data sharing, will pass these habits on to students. This concern is valid, especially because it has been found that a researcher's attitudes to the sharing of data have a significant effect on student behaviors.[7]

We understand that many of our problems in data management result from cultural rather than technical issues. Mark Parsons from the Research Data Alliance sums this issue up when he speaks about the "long tail of data" or the vast amount data that has been created by individual researchers or small research groups "To have some sort of consistency across that data so that we can integrate it requires a lot of social change."[8]

Combined effort needed

Institutions should be prepared for these changes and understand that researchers soon will require increased support. Researchers and stakeholders who are aware of changes also should be prepared to explain them to those who

are not aware, including administrators, IT staff, librarians, sponsored research officers, and others. Without a combined effort, individual researchers will struggle to meet requirements, especially as agencies continue to execute further stages. Institutions that do not begin to adopt data management planning and best practices will find themselves behind, which may affect funding and recruitment of students and faculty.

The materials field

Materials science researchers have talked about the topic for years (or at least talked about issues that have resulted because of the lack of good practices) without using the words "research data management." Stephen Freiman and John Rumble have written on this topic describing the current state of data and challenges facing the materials field. They have covered the description of nanomaterials, changing nature of materials, access to proprietary data, and, in general, uncertainty of documentation and data reliability. This sentiment was echoed in a report funded by the Department of Defense, which acknowledged that access to ceramic property data was "haphazard" and that much of the data available lacked provenance and quality indicators.[9] A more recent study on materials science and engineering data describes the current state as lacking "the strategy, framework, standards, and culture needed to support materials data curation and sharing."[10]

Materials science, similar to academia, is struggling with data management. Although there is an obvious need for developments, such as database linking, data mining, and single-point access, the data going into these systems first must have been well-managed through the research cycle. If we expect to make progress in materials data, data management and data sharing must be stressed. Education is key to success, but stakeholders in the field must take a step back to examine the culture, ask questions, and work to understand why this topic has gone unaddressed for so long.

Considerations

Many issues must be considered as the community of researchers, administrators, and technical and information professionals work toward development of support systems and standardized practices. Although the scope of this article is limited, the following are examples of tough questions that should encourage thought and conversation. There is no "solution" to research data management as a whole—rather it is the importance of making incremental improvements. These questions should make stakeholders think how solutions may affect their research or the support provided to researchers.

- Why would tenure-track faculty members (or any faculty member) focus time and effort on curation of their data for purposes of sharing when they will be judged only on publications?
- As researchers look toward retirement, who is responsible to gather and preserve important data that has not been shared? How can this data be used to teach students on best practices?
- What incentive do graduate students have to deposit research data along with their theses? How could this requirement impact their research?
- How can funding agencies encourage researchers to work with their institution's IT staff, librarians, and administrators to build a network of support?
- What value does a graduate with a good understanding of data management bring to industry? Is the return on investment great enough for companies with robust practices to work with universities on skill development?
- What is the cost of not integrating data management into education?

Progress

Astronomy seems to be a good example when we look broadly for fields that have made the most progress. Kevin Ashley, director of the Digital Curation Centre (DCC, Edinburgh, Scotland), spoke about the changes within astronomy recently at the University of Warsaw Center for Open Science.[11] He explained that, within the field of astronomy, a major catalyst to change was the shift in the way data were captured. Once the capture of data moved from photographic films and plates to a digital form, sharing became simpler. Technology had forced the field to make decisions on issues surrounding data management and sharing. The field adopted and enforces a six-month exclusivity (in most cases) on data. After that, data is available to all researchers. In the interview, Ashley explained that this practice has enabled better research, and more publications are coming out today that are based on second-hand observations than on original observations.

Curriculum development

As information professionals, librarians are key players in data management. Because of their skillsets and background, librarians are often relied upon to help manage research data. This involvement is evident in the body of literature on data management and in the educational resources that have been developed. Two examples of this are the New England Collaborative Data Management Curriculum (NECDMC) and the Data Information Literacy Project.

The NECDMC comprises seven modules that align with NSF's data management plan recommendations. It also addresses challenges that face researchers in the sciences. The curriculum uses case studies for context and even offers documentation on how to teach courses using the material provided. The National Library of Medicine funds the project, which is available free for use under a Creative Commons (CC-BY) license.

The Data Information Literacy Project is a collaboration by the libraries at Purdue University, Cornell University, University of Minnesota, and University of Oregon. The project's goal is to develop an infrastructure in which students are taught data information literacy (DIL) skills relevant to their discipline,

while helping to develop a process in which DIL curricula can be articulated within research communities. The most relevant outputs thus far include case studies at each university and a data management course. Both are available online and help lead students through the development of data management plans. The Institute of Museum and Library Science funded this project, and course materials also are available free under a CC-BY license.

Within the College of Ceramics at Alfred University, the Scholes Library staff leads the conversation on research data management. The library has developed graduate seminars, integrated data management planning in coursework, and embedded services to ensure that students and faculty understand new requirements and are better equipped to meet them. A smaller support and technical infrastructure means that services cannot be developed immediately. However, the library has taken advantage of its SUNY connection in planning a path forward.

Contributing to a solution

There are ways that government, industry, and academia can contribute to improving data management practices. There are organizations that provide services related to research data management, including Research Data Alliance, University of California Curation Center, and Digital Curation Centre. The role of these organizations in developing data management is important. More specialized organizations, such as The American Ceramic Society, are good places to facilitate conversation. Because ACerS has members in all sectors, researchers and other stakeholders can rely on a vast pool of expertise. Development of a data management framework in materials requires the input and knowledge of various groups.

If the idea of data management, whether in our own research or on a larger scale, seems daunting, we know we are not alone. In the end, however, we cannot make progress without involvement by individuals. So what can we do?

- *Educate ourselves:* Better understand best practices, look at case studies, and work to better understand new requirements.
- *Communicate:* Talk with colleagues, students, professors, and supervisors and work to understand how others are managing data within our institutions.
- *Advocate:* Work with our institutions' professional staffs to develop data management training options for undergraduate and graduate students, postdoctoral researchers, and research support staff.
- *Evaluate:* Make changes to the way we manage data and work with others to improve data management practices at a local or group level.
- *Lead:* Reach out and work with others to form a group within an organization (such as an ACerS Technical Interest Group).

About the author

Trevor Riley is engineering and emerging technologies librarian at Alfred University's Scholes Library. Contact Riley at riley@alfred.edu.

References

[1]"Require free access over the Internet to scientific journal articles arising from taxpayer-funded research," U.S. White House, 2012, accessed July 2015, available at <https://petitions.whitehouse.gov/petition/require-free-access-over-internet-scientific-journal-articles-arising-taxpayer-funded>

[2]"Increasing access to the results of federally funded scientific research," Office of Science and Technology Policy, 2013, accessed June 2015, available at <https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf>

[3]"Today's data, tomorrow's discoveries," National Science Foundation, 2015, accessed June 2015, available at <http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>

[4]J. Carlson and M. Stowell-Bracke, "Data management and sharing from the perspective of graduate students and examination of the culture and practice at the water quality field station," portal: *Libraries and the Academy*, **13** [4] 343–61 (2013).

[5]"Research data management: Principles, practices, and prospects," Council on Library and Information Resources, 2013, accessed June 2015, available at <http://www.clir.org/pubs/reports/pub160/pub160.pdf>

[6]A. Diekema, A Wesolek, and C. D. Walters, "The NSF/NIH effect: Surveying the effect of data management requirements on faculty, sponsored programs, and institutional repositories," *The Journal of Academic Librarianship*, **40**, 322–31 (2014).

[7]Y. Kim and P. Zhang, "Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories," *Library and Information Science Research*, **37** [3] 189–200 (2015).

[8]"Openness should be the default for all research data: Maciej Chojnowski in conversation with Mark Parsons," Interdisciplinary Centre for Mathematical and Computational Modeling at the University of Warsaw, 2015, accessed September 2015, available at <https://www.youtube.com/watch?v=3huPI2DJIEQ>

[9]"E-Ceramics 2012: The availability, quality, and location of ceramic property data," Department of Defense, 2012, accessed July 2015, available at <https://ceramics.org/wp-content/uploads/2013/03/00-ceramic-data-report.pdf>

[10]C.H. Ward, J.A. Warren, and R.J. Hanisch, "Making materials science and engineering data more valuable research products," *Integrating Materials and Manufacturing Innovation*, **3** [22] (2014).

[11]"Benefits for society comes when you open the data up and allow its exploration: Maciej Chojnowski in conversation with Dr. Kevin Ashley," Interdisciplinary Centre for Mathematical and Computational Modeling at the University of Warsaw, 2015, accessed September 2015, available at <https://www.youtube.com/watch?v=6H_Eu5r23ec >

More data management @ ceramics.org

- White House calls for increased access to federally funded research results **ceramics.org/white-house-calls**
- NIST issues report on MGI workshop addressing data and standards **ceramics.org/nist-issues-report**
- Data drives engineering of ceramics; workshop asks 'how well?' **ceramics.org/data-drives-engineering-of-ceramics**

Copyright of American Ceramic Society Bulletin is the property of American Ceramic Society and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.