CASC Coalition for Academic Scientific Computation

Research Data Management Position Paper

October 2023

Research data has grown explosively over the past couple of decades, and the research community is at a critical point for developing broad standards for managing data from the research project ideation phase through post project long-term archiving. Funding agencies have introduced requirements for data management plans, data accessibility, and data retention, but those requirements often come with no associated funding or a tradeoff between funding research activities or data management activities. There are several important questions facing our institutions and our communities, including: Who owns research data? Who is the steward for research data at various stages of its lifecycle? How is research data management funded? Who decides when to keep data and when to get rid of data? Who is responsible for creating and maintaining metadata that will enable automation for management processes? This Research Data Management Position Paper represents a first step in enumerating the challenges these emerging requirements represent and presenting a series of recommendations directed at CASC, federal sponsors and the RDM community itself to support and advance the principles of open, accessible science.

Key Takeaways

- Data management and storage has been a looming challenge at research institutions as technology has made it possible to generate ever-larger datasets to explore an ever-wider array of research questions. Storage and management of data have largely been the domain of individual researchers. Institution-level storage and management often fall as “make-do” expectations on under-resourced support services in information technology departments, libraries and offices of sponsored research. In centuries past, data management was the responsibility of the libraries, and institutional funds were provided for that service. The cost of that service has been included in indirect cost negotiations with the federal government. However, the explosive scale of digital data (research data, in particular) far outpaces the responsibility of libraries alone and will require a cross-disciplinary scope that addresses the challenges from a technology, domain-level, and governance framework. That situation needs to be updated at the highest levels in government and at institutions.
- As federal sponsors adopt policies that call for data management and storage plans that support principles of open science, CASC should use its position to advocate for clarity, consistency and funding, which encompasses both the technology and the personnel needed to comply.
- CASC, in turn, should help members demonstrate the value proposition of institutional investment in cyberinfrastructure—hardware, software and

personnel — as a means of reducing risk, attracting talent and enabling discovery.
- CASC has the opportunity to help shape this movement toward more open and accessible research by contributing to the development of common practices and a template for RDM architecture, as well as helping to refine the research life cycle model to reflect the data management needs of today and the future.

Emerging Data Management Expectations Raise Concerns

In August 2022, the Office of Science and Technology Policy instructed leaders of federal sponsors of research to update their policies on public access to data. The National Institutes of Health was the first agency to act, implementing a Data Management and Sharing (DMS) policy in January 2023 to set forth expectations of investigators and institutions pursuing funding. Other federal sponsors of research, including the National Science Foundation, are expected to take similar steps to promote sharing of scientific data. The OSTP memo set a deadline of no later than December 31, 2025, for federal agencies to update policies on access to research.

The NIH policy, which requires grant seekers to submit a plan for managing and sharing data and threatens those who fail to comply with DMS plans with the loss of funding, has already generated considerable concern and uncertainty within the research community. Chief among these are:

A lack of clarity: The growing focus on storing and sharing data in support of making research accessible and reproducible raises a central question: What data? Computational modeling involves huge amounts of data, but it is unclear what data are necessary to be in compliance. Raw data? Processed data? Results of analysis? All data used for a funded project, even that not included in articles that arise from the research? It can be difficult to determine what data are required to be shared, as well as the software necessary, to meet the expectation of reproducible research. This can be especially difficult when there can be many collaborators on a project, as well as multiple institutions. As the volume of data escalates, this lack of clarity poses questions of maintaining adequate capacity, advising researchers regarding consistent format for storage, protecting private and sensitive information, and ensuring integrity of data and models from inadvertent user errors and potential "bad actors." This problem calls out the need for the creation and maintenance of metadata, which, through automation, is the only way to manage data at scale.

A lack of funding: One central challenge to achieving the goal of data storing and sharing requirements issued by NIH and expected by other federal sponsors is infrastructure build-out. Institutions lack the resources to invest in the cyberinfrastructure necessary to store the amount of data that will result. Institutional administration outside of the research data management community may tout that they offer every faculty member gigabytes of storage, which may have been sufficient at one time, but achieving the principles of data accessibility and reproducibility will require an exponential shift to terabytes, petabytes and even exabytes. Higher resolution images, such as occurs with MRIs, metal purity samples and many other common applications, result in data files that are relatively cheap to generate but still relatively costly to store, especially in a way that promotes ease of sharing. The NIH guidelines expect grant seekers to anticipate the cost of maintaining

project-related data for three years, but this represents a piecemeal approach to institutional buildout of sufficient cyberinfrastructure. Well-resourced universities may be able to invest upfront in anticipating the storage needs of their researchers, but less-resourced universities will struggle to stay ahead of capacity needs.

Risks to privacy and security: Research for federal agencies such as the NIH and Department of Defense come with expectations of privacy and security, complicating the principle of open science. Making research involving private health data accessible and reproducible requires more than usual steps taken to deidentify data. The NIH offers specific policies involving human genome research, but medical imaging research and health studies examining traditionally marginalized groups are two other examples where privacy concerns potentially conflict with principles of accessibility. Information about race, age, income, and zip code may be sufficient to reveal personal health data of deidentified research subjects. Conflicting expectations of open but restricted access adds a level of complexity — and risk — to institutions providing campuswide repositories for federally sponsored research.

Similar to privacy concerns regarding health and other personal information, research with federal sponsors such as the departments of Defense and Energy involve restrictions on who has authority to access information. The National Institutes of Standards and Technology has put forth an updated set of compliance standards (NIST 800-171) governing ways in which universities and nonfederal systems must ensure the security of sensitive information they collect and store. Ensuring a level of security sufficient for research sponsored by Defense and Energy departments requires costly investments in hardware, software and staffing. In addition, compliance requirements may impact research workflows, even activities that do not fall under compliance expectations, to the point that they jeopardize the research mission of the university. Federal expectations of being both accessible and secure add even more complexity, which may not be fully understood at the level of university administration charged with signing off on grant proposals and data management and storage plans. Institutions where medical and defense research is common may have the systems and sophistication to navigate potentially conflicting expectations, but most institutions engage in a wide mix of research activities, making maintaining privacy while providing accessibility a cause for concern. No institution wants to find itself in the news for a security breach, but university leadership often fail to appreciate fully the risks associated with insufficient cyberinfrastructure and data management protocols and inadequately trained support personnel.

No systematic support: The NIH Data Management and Sharing Policy sets expectations for grant applicants regarding “identifying appropriate methods/approaches and repositories” and allows for requesting funding for such activities—unless data management and sharing services are already provided by the institution or some other source. The policy makes clear that individual researchers maintaining their own data sets on their own equipment does rise to the envisioned accessibility to high-value data sets that enables validation of results and accelerates discovery. It sets an expectation that is likely beyond the ability of individual applicants, particularly those at less-resourced institutions, supports piecemeal funding for what likely needs to be a more broad-based data management solution, and appears to penalize institutions that have already invested in sophisticated cyberinfrastructure. Moreover, it fails to account for what happens to data beyond its useful life. Who is responsible for the cost of monitoring,

making available and potentially purging data beyond the storage period?
No standard approach: As federal sponsors move toward prioritizing and requiring accessible and reproducible research, institutions may benefit from sharing their own experiences in implementing campuswide plans for data management and storage. Yet, even those who are further along in their journey, vary widely in their approaches and in designating who takes the lead. Responsibility for implementing a plan for RDM typically is shared across three campus departments: information technology and computer centers, libraries, and offices of sponsored research. Universities vary in terms of which department takes the lead and the extent to which each contributes. It may be too soon in the movement to identify "best" practices, but an effort to collect and assess common practices could help bring clarity and standards that will enable a smoother, more secure transition.

A potential to exclude: The stated goal of the NIH policy is to make science more accessible and more innovative, but prioritizing data management and storage among criteria for proposal funding may present a barrier for under-resourced universities, including minority-serving institutions. The requirement favors grant applicants at institutions with the capacity for "team science," meaning they have the existing infrastructure and staff to support RDM. Although grant seekers are able to request funds for data storage and specialist support into proposals, that may take focus — and potentially funding — away from core research activities. It also fails to account for the need for universities to make initial investments in personnel and storage capacity and to develop a sustainability plan to cover those costs over time.

Steps Toward Addressing Concerns and Achieving Open Science Vision

Actions for the RDM Community

CASC occupies a pivotal position in supporting and speeding widespread adoption of open science principles. CASC should use its position as advocate and adviser to help shape policies adopted by federal sponsors and to disseminate information and lessons learned with its members. Although CASC currently cannot receive grants, the ability to shape this open science transition demonstrates the opportunity for CASC-led supported research. With adequate funding, CASC could survey its members on common obstacles and common practices, as well as providing input to help answer questions of suitable formats for saving and uploading data, data value and life cycles, and ethical and equitable access.

In the near term, CASC should share with federal research sponsors the concerns of members regarding clarity, funding and incentives. CASC should amplify member concerns regarding siloed efforts of federal agencies that create a patchwork of RDM regulations. Many CASC members are tasked with supporting many different departments and activities at their institutions. Clarity is needed from federal sponsors regarding identifying how much data to save to support goals of reproducing research, valuing the potential impact of data, navigating the challenge of being open but secure, evaluating when it is cheaper to save work versus recreating anew, and developing consistent plans, using automated tools that work at scale and are driven by metadata, for eliminating data that have exceeded their expected life cycle.

As a longer-range goal, CASC should leverage the expertise and experience of members to develop a template research data management and storage plan. Drawing on common practices, the template would both help to standardize RDM

protocols and systems, while also providing an “off-the-shelf” solution for institutions less equipped to produce accessible, scalable and secure cyberinfrastructure plans. As CASC explores how to better represent the broadening research computing ecosystem, the RDM template, as well as other support in the form of mentoring or information sessions, would demonstrate the value of membership to a wider mix of institutions.

CASC could also demonstrate its value to federal sponsors and the research data community by convening work sessions with the goal of refining the existing research data life cycle model to be more tangible and relevant in reflecting the shift from a data storage process away from individual researchers to an institutional process.

Those who support the research data ecosystem must also advocate for themselves on their own campuses. They should look for ways to break down institutional silos and build relationships with researchers as well as administrators. Instead of focusing on costs, they should advocate the value of research data management as enabling discovery, reducing institutional risk and attracting talent. They also need to engage with counterparts at other institutions to exchange ideas, envision needs, and shape the expanding research data community of the future.

Actions for Federal Sponsors

This position paper stands as a direct appeal to federal funders on behalf of the research data community. Federal sponsors of research need to come together and agree upon a standard approach to supporting principles of open science. One policy covering all agencies or similar policies across agencies that set forth shared expectations would enable better compliance on the part of researchers and institutions.

Specifically, federal sponsors need to provide institutions with better guidance regarding the cyberinfrastructure and support expected to achieve adequate retention, sharing and security. Carve-outs may be warranted in certain circumstances, but shared expectations that cut across funding priorities would enable institutions to implement campuswide RDM strategies and systems that smooth the adoption of open science principles and practices.

Federal sponsors would do well to remember the adage about getting more of what is incentivized. The NIH’s policy purports to value accessibility and reproducibility of data, but funding priorities and criteria still tend to incentivize new research over that which reuses and reproduces. Federal sponsors should assess how funding decisions align with and support stated goals.

As federal agencies enact policies in pursuit of open science principles, they should consider whether relying on individual institutions as repositories best achieves the NIH’s stated goal of accelerating the pace of research and discovery. National repositories for funded research may better support accessibility and ensure security. Alternatively, regional repositories may allow minority, rural and other less-resourced universities to pursue federal support for research without facing the challenge of developing and maintaining their own RDM plans. A regional approach may also help to address a likely shortage of workers with the skills and expertise to operate and maintain sophisticated cyberinfrastructure systems. Institutions and regions vary in their ability to develop and attract talent so RDM expectations as a criterion may represent an exclusionary burden.

Authors: Fran Stewart, Joe Hetrick, Jeremy Frumkin, Brian Hammond, Joshua Baller, Kim Wong, Jackie Milhans, Sarvani Chadalapaka, Tabitha Samuel, Kaylea Nelson, Birali Runesha, Robert Bjornson, Hongfeng Yu, Vikram Gazula, Thomas Cheatham, Jarek Nabrzyski, Mike Warfe, Erik Deumens, Kathryn Kelley, Carolyn Casler
Executive Committee: Dave Hart, Rich Knepper, Barr von Oehsen, Jim Wilgenbusch