# Public Access to Research Data Working Group Report

Submitted on November 11, 2019


Jake Carlson, Director of Deep Blue Repository and Research Data Services, Library (Chair)
Sol Bermann, Executive Director of Information Assurance and Chief Information Security Officer, ITS
Lois Brako, Assistant Vice President for Research Compliance, UMOR
Jason Garr, Associate General Counsel / OTT, OVP and General Counsel
Sharon Broude Geva, Director of Advanced Research Computing, UMOR
Erin Kaleba, Director of the Data Office: Clinical and Translational Research, Med. School
Myra Kim, Research Scientist, CSCAR
Maya Kobersy, Associate General Counsel, OVP and General Counsel
Margaret Levenstein, Director, ICPSR
Asmat Noori, Information Systems Security Manager, ITS
Tracy Pattok, Associate Director for Institutional Research, OBP
Craig Reynolds, Executive Director, Office of Research & Sponsored Projects, ORSP
Nick Wigginton, Assistant Vice President and Chief of Staff, UMOR
Diane Wilson, Regulatory Affairs Manager, Med. School


**Executive Summary:**

Universities are under growing pressure to provide support to researchers in managing, sharing, and preserving the data they generate. Funding agencies, publishers, and other external stakeholders have instituted requirements that research data be made publicly available to those seeking to better understand the research process, to replicate or reproduce the research, or to re-use the data that was collected to conduct new research. These requirements were introduced as the result of efforts by forward thinking researchers and the general public to open scientific inquiry, build trust in research, and promote good research practices.

The University of Michigan (U-M) has responded with resources and services to help researchers fulfill their data sharing obligations; however, much of the data generated at U-M remains largely inaccessible. Furthermore, in a recent survey, a high percentage of principal investigators at U-M reported that their data management responsibilities constitute a substantial workload. Moreover, the scope and extent of data sharing requirements from funding agencies and expectations for access to research data continue to deepen and evolve. To retain our considerable advantage in attracting research funding and to preserve our research reputation, U-M needs to become a leader in making data publicly accessible. We should take on this leadership role, not just to respond to ever-increasing expectations for access to data, but to anticipate and exceed them.

In assessing the current environment and the support U-M provides to researchers we see three broad and interrelated challenges. First, there are currently few incentives for researchers to invest their scarce time, energy and resources in sharing their data. Fostering a climate in which publishing data is recognized and rewarded as a scholarly achievement, particularly in making promotion and tenure decisions, would shift the perception of data sharing as solely a burden. Second, the scale and heterogeneity of the data produced by researchers at U-M make it difficult to provide both the amount and specificity of support needed to individual researchers. In addition to considering where additional investments may be needed within U-M to expand our capacity and capabilities, we will need to forge cross-institutional partnerships to create more data sharing networks. Third, the lack of a common understanding of data sharing requirements, and a lack of coordination in providing services, increase the burden to researchers and artificially limit U-M's capabilities to share data of value. We need to define responsibilities for supporting data sharing more clearly and integrate this work across support units to increase our effectiveness and efficiency in making data accessible.

Solving the challenges of providing public access to research data generated at U-M will not be done overnight, but U-M has a strong foundation from which we can build. Our mission statement recognizes that U-M serves the people of Michigan and the world through creating, communicating, preserving, and applying knowledge. We believe that enabling public access to research data is integral to this mission.

The working group developed the following recommendations to begin addressing the challenges we identified:

1. U-M should create a Committee on Data Sharing comprised of faculty from across the university and charge them with acting on the issues identified in this report. Researchers at U-M have interpreted and responded to data sharing requirements and issues individually. Creating a university Committee on Data Sharing will enable a more coordinated and unified approach to making investments, addressing common issues, and meeting requirements.

   We see three initial areas of responsibility for the committee to address.

   First, this committee should take steps to develop an understanding of researcher experiences across U-M in responding to data sharing requirements, and how they make use of (or do not make use of) the services and support provided by U-M. Developing a broad cross-institutional understanding will inform where and how U-M should invest resources to ease the burden of data sharing and maximize its potential impact.

   Second, this committee should apply what they learn towards developing and promoting a shared set of institutional values at U-M around data sharing. These values would provide a strong foundation to build connections across the university and coordinate needed services, and support for sharing data.

   Third, this committee should address the gaps and inadequacies of U-M's policies as they pertain to managing, sharing, curating and preserving research data. An important part of this effort will be to reexamine how roles and responsibilities on research data are defined, distributed and coordinated.

2. We propose forging a closer relationship between units through creating a Research Data Sharing Services Group (RDSSG). No single unit on campus can address all of the many complexities of data sharing and so a more coordinated approach is needed. The RDSSG would strengthen relationships, bolster communication, and foster collaboration for supporting data sharing at U-M. This group would also work closely with the faculty Committee on Data Sharing to address issues in administering data sharing requirements and service provision, including:

   - Taking steps to better account for the data produced at U-M;
   - Incorporating data management plans more directly into the University's grant support systems;
   - Reviewing the costs and allocations of digital storage in support of making data publicly accessible over time;
   - Developing educational programming and encouraging departments to include data sharing into their curricula;
   - Adopting persistent identifiers to aid in the discovery and attribution of shared data.

**Introduction:**
Public access to data is increasingly a critical component of 21st-century scholarship. In 2013, the Office of Science and Technology Policy released a Memorandum directing the Heads of Executive Departments and Agencies to take steps to ensure that peer-reviewed papers and research data are made available for public use. This wide-ranging Memorandum outlined a vision for public access to research data that included developing strategies for making data findable and accessible while being securely stewarded to enable long-term use. As a result, federal agencies with a budget greater than $100 million were tasked with developing plans for ensuring public access to all outputs of federally funded scientific research. Many private funding agencies have followed suit in requiring that researchers share the data generated from their work with the public as a condition of making the award. Publishers and learned societies increasingly require that data used to support the research findings described in an article be available for readers to access and review.

It may be tempting to see the directives for public access to research data coming from the government, funding agencies, and publishers as top-down, bureaucratic driven, unfunded mandates. In fact, however, the 2013 Memorandum and other similar initiatives are the outcomes of efforts by multiple research communities and the public to make data more widely available for use outside of their original context. Researchers in genomics, astronomy, and social science fields, for example, have demonstrated the benefits of having access to rich, high-quality data to support their research.

Universities have important roles to play in supporting public access to research data, but providing this support effectively requires that university administrators and service providers address an inherent tension in how research is practiced. Researchers are part of larger communities comprised of others engaged in their field of study. The expectations, norms, and standards of practice in conducting and disseminating research are developed and supported by these communities. Funding agencies and journals serve as extensions of these communities by providing support for undertaking and communicating research. The *work* of research, however, generally takes place at a university or similar institution. These institutions are tasked with providing infrastructure, staff, and other resources for multiple fields that employ a variety of different research methodologies and practices, each of which require specific types of services and support. Though researchers are the ones submitting proposals to funding agencies, the awards are generally made to the university itself. Ultimately, then, it is the university that is responsible for fulfilling the requirements of the award. Although the institution plays a vital role in providing services for researchers, including supporting the management, dissemination, and preservation of research data, discussions around facilitating public access to research data have tended to overlook the role of the institution in favor of discipline-specific communities and funding agencies.

Fostering an environment in which researchers can easily make their data accessible to the public is in the best interests of U-M. In addition to supporting individual researchers who need external funding to support their work, there are multiple benefits that come from making data publicly accessible.

- Data sharing strengthens open scientific inquiry, good research practice, and the reproducibility of research.
- Having the data readily accessible along with journal articles and other outputs helps the audience better understand the findings of the research.
- Sharing data promotes collaboration amongst researchers and reduces costs through avoiding the duplication of data collection efforts.
- Making data available to others encourages better management and documentation of the data, leading to higher quality outputs. It discourages falsifying data and academic fraud.
- Multiple researchers are able to perform their analyses using the same data, enabling a richer academic discourse.
- Transparency and sharing data from research paid for by tax dollars promotes public trust in the research enterprise and in academic institutions.
- Studies indicate that sharing data results in a greater impact for the research project and higher profiles of the researchers themselves.

Furthermore, sharing data is closely linked to the mission of U-M. In our mission we state that the University has a responsibility to serve the people of Michigan and the world through communicating and preserving knowledge. Increasingly, research data are an integral component of the knowledge that is created at the University. We have taken on the responsibility to lead in developing and living the academic values that will challenge the present and enrich the future. Public access to research data is rapidly emerging as a critical value in supporting research in the 21st century and beyond.

In November 2017, the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) released a report[1] with recommendations for universities and federal agencies to adopt to ensure public access to federally sponsored research data. This report was followed by a workshop held in October 2018 for representatives of universities to gather and discuss the role of the university in supporting public access to research data. U-M sent a team of delegates -- representing the Office of the Provost, the Office of Research, the University's General Counsel, the School of Information, and the Library -- to the AAU-APLU workshop. Attending the workshop further reinforced our sense that the expectations of accountability being placed on institutions for ensuring public access to research data will continue to grow. The actions, or inaction, of U-M to these increasing pressures will almost certainly affect our ability to secure the resources we need to achieve our mission.

The Provost and Vice President for Research have recognized that funding agencies and publishers are serious about their requirements for making research data publicly accessible. In response, they charged this working group to study the issues surrounding public access to research data and make recommendations on how U-M could improve compliance, reduce the burden to researchers in sharing their data, and make the process as seamless as possible in a fiscally responsible manner. The University recognizes the

[1] AAU-APLU Public Access Working Group Report and Recommendation. November 29, 2017 https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations

importance of data sharing to science, because it supports reproducibility and leverages prior investment for new knowledge creation. This report is the primary deliverable of the working group.

**Definitions**
Public access to research data could conceivably be defined in multiple ways. We recommend that U-M adopt the following definitions:

- *Research Data* is defined as the recorded factual material commonly accepted in a scholarly community as necessary to validate research findings. Research data may be defined by its context (ex. administrative data can potentially be used for research purposes) or use (e.g., analysis of tweets or newspaper text for research purposes). Depending on the type of research performed, data could include software, physical specimens, experimental protocols, or other materials.

- *Public Access* is defined as having a transparent process for the public to gain access, or request access as appropriate, to research data generated at U-M.

It is important for us to state clearly that public access to research data does not mean that everyone would be able to access and use any specific data set for any reason. For example, when data sets contain sensitive information, the University has a responsibility to establish the means and processes for reviewing requests for access and to permit them only as warranted after being vetted. Safeguards and checkpoints must be put into place to protect confidentiality of research subjects as well as intellectual and commercial property rights of researchers and the university.

- *Data Management* is defined as the actions taken by a researcher or other persons with responsibilities for the data to ensure that the data are fit for use, secured, and protected from harm as the data are actively being developed.

- *Data Curation* is defined as the actions taken by the researcher or a third-party to add value to data and increase its utility and usefulness to others outside of the environment in which the data were developed. Curation often includes activities taken to enhance the organization, description, accessibility, and preservation of the data.

- *Data Preservation* is defined as the actions taken by an information professional, such as an archivist or librarian, to lengthen the lifespan of the data and mitigate against its deterioration, loss, or obsolescence. An important element of data preservation is determining which data sets have sufficient value to merit their preservation and for what duration of time. Not all data should (or could) be preserved; nor should it be expected that data will be preserved forever.

**Challenges in Making Research Data Publicly Accessible**

Research data are not created fully formed; rather they are developed and shaped over time and in stages. Collectively these stages can be referred to as the data life cycle. The specific stages that a particular data set will pass through vary depending upon the nature of the research being conducted, the norms and expectations of the field of study, and the preferences of the researchers themselves, among other things. However, there are broad stages that commonly occur in the life cycle of data, as depicted in the figure below.

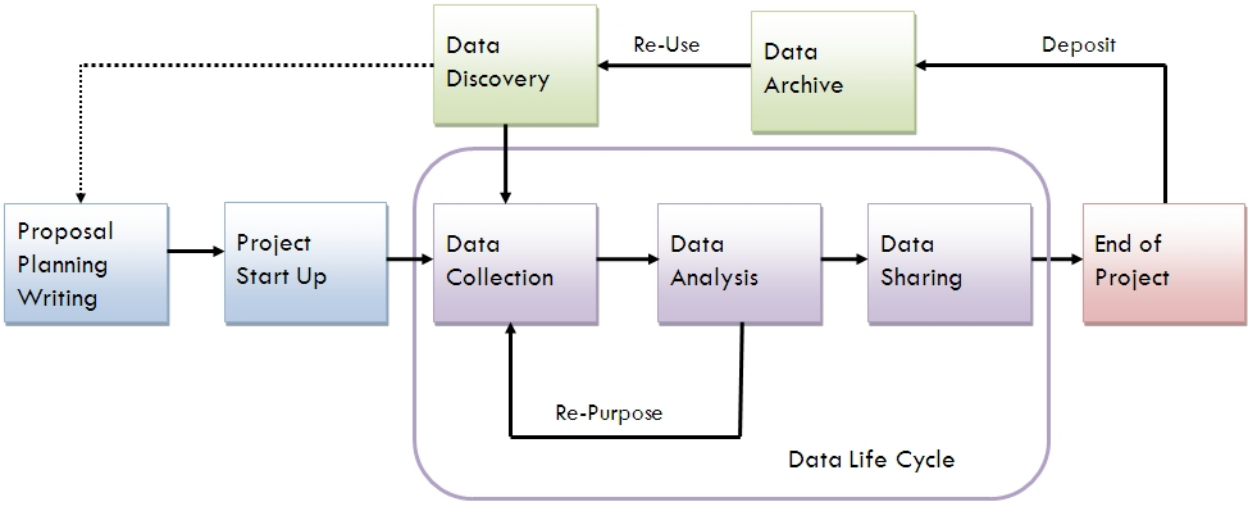


Figure 1 - An example of a generic Data Life Cycle

The "Proposal Planning Writing" and "Project Start Up" stages occur in the initial or planning phase of the research project. It is here when funding for the research is sought, data management or sharing plans are written, and Responsible Conduct of Research or other training is delivered. The data and the processes used to develop and protect them are described and documented in this stage.

The "Data Collection," "Data Analysis," and "Data Sharing" stages occur when research is actively conducted and plans for data development are enacted. In these stages, data are first collected or acquired, processed to prepare them for analysis, and analyzed to develop answers to the research questions being posed. Data may be shared to varying extents with colleagues or within the research outputs as charts, tables, or graphs. Data typically require secure storage, the creation of back-up copies to protect against disasters, periodic quality assurance checks, and the generation of documentation and metadata.

The "End of Project" stage represents the myriad of activities that take place when a research project reaches the end of its active life span. At this point attention turns to closing out the project and ensuring that all requirements and obligations were met. This includes implementing the commitments for sharing and preserving the data that were made in the data management plan.

The "Data Archive" and "Data Discovery" stages depict the data being deposited into a data repository or otherwise being made publicly available by the researchers. At this point the

data are findable, accessible, interoperable, and reusable by others outside of the individual or team who developed the data. Ideally, the data are actively being curated and preserved by a repository to make the data more useful to potential consumers.

Equally as important as the stages themselves are the connections between the stages, represented in the figure by the directional arrows. Each stage in the data life cycle is informed by or dependent on what takes place in the previous stages. For example, ideas for research projects are fueled by access to data and other sources of information. A well-defined, properly resourced, and actionable data management plan supports the researchers' ability to successfully manage and work with their data at each stage in their data's life cycle. Taking steps to share data early in the data life cycle reduces the cost, and improves the resulting data quality, later. Perhaps most critically, depositing data into a data repository or archive requires that the researcher prepare the data earlier, during the stages when it is actively being developed. Transferring data from active use storage to a publicly accessible repository runs the risk of breaking significant connections between components and losing important contextual information. Without preparation, the data become more difficult for others to discover, understand, trust, or use. Even with good planning and data management the process of depositing data into a repository can be challenging. This is particularly true for researchers in fields that have not yet developed standardized tools, resources, and practices around sharing and preserving data.

In addition, simply putting research data online is insufficient to comply with federal mandates or achieve the benefits of making data publicly accessible. Data need to be accompanied by documentation and metadata that thoroughly describe the content of the data and how it was developed. Documentation can be defined as the human readable information needed to understand, trust, and reuse the data. Metadata is the contextual information needed for people to be able to find the data through an internet search engine such as Google, or a data repository or catalogue. Metadata should also provide enough contextual information to enable someone to evaluate the data set well enough to decide if it is of interest or use to them. Metadata should be readable by humans and actionable by machines. The amount, format, and content of the documentation and metadata needed will vary by data set and the expectations of the researcher's field of study.

**Challenges and Risks to the University of Michigan**

U-M is a world-renowned public research institution with annual research expenditures of more than $1.5 billion. In FY18, approximately $820M (or 96%) of U-M's $852M in federally funded research expenditures were subject to the data sharing requirements of one of the sixteen federal agencies to have such a requirement. In that vein, the top five funders of U-M research -- all of which have a data sharing requirement -- were: (1) the National Institutes of Health (NIH), $538M; (2) the National Science Foundation (NSF), $83M; (3) the U.S. Department of Defense (DoD), $79M; (4) the U.S. Department of Energy (DOE), $41M: and (5) the National Aeronautics and Space Administration (NASA), $29M.

The federal agency expectation of grantees that they share their research data is not new. The NIH, U-M's largest source of federal funding, has required since October 2003 that applicants seeking $500,000 or more in direct costs in any year of the proposed project

period include a data sharing plan in their proposal. The NSF, U-M's second largest source of federal funding, has since January 2011 required that every applicant include a two-page supplementary document consisting of either a data management plan (DMP) or an explanation of the absence of the need for such a plan. The components of a DMP include dissemination, access, and sharing of data; policies or provisions for re-use, re-distribution, and production of derivatives from the data; and archiving of data.

The data sharing requirements of funding agencies continue to evolve, and agencies are stepping up their enforcement of data management or sharing plans. The NSF's treatment of their data management plan requirement is a case study in how data sharing requirements across funding agencies are taking shape. Even before the DMP requirement was instituted in 2011, the NSF had recommended data sharing for grant recipients since 2005. Review of the DMP requirement was limited when it was introduced but reports from program officers and reviewers indicate that the content of DMPs are increasingly scrutinized. NSF's 2015 report on public access to data states that DMPs are reviewed "as an integral part of the merit review of the proposal, considered under Intellectual Merit or Broader Impacts or both."[2] Each of NSF's directorates continues to update and revise guidelines around the DMP to strengthen the requirements and their enforcement. The NSF's Engineering Directorate, for example, revised their guidance in 2018, adding a requirement that: *"The PI(s) must manage their data as described in the DMP and will be monitored primarily through the normal Annual and Final Report process and through evaluation of subsequent proposals.[3]*" The NSF itself continues to consider and develop expectations around data sharing as well, as evidenced by its May 2019 "Dear Colleague Letter," in which specific encouraged practices were called out: persistent IDs for research data and machine-readable DMPs.[4]

Despite ongoing developments around strengthening the data sharing requirements of funding agencies and their enforcement, there has been little guidance to PIs as to the expectations for successful adherence. A recent report on faculty workload at U-M highlights the challenges institutions face in ensuring compliance with data sharing requirements.[5] The report presented the results of a survey given at U-M in 2018 (n=300 faculty), which included a section on data management responsibilities. Survey results indicated that 75% of researcher respondents at U-M had data management responsibilities and that 45% of them considered their data management responsibilities to be a substantial workload. Fifteen percent of survey respondents saw data management

---

[2] Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation. National Science Foundation. March 18, 2015
https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

[3] Directorate for Engineering Data Management Plans Guidance for Principal Investigators. November 2018.
https://nsf.gov/eng/general/ENG_DMP_Policy.pdf.

[4] Dear Colleague Letter: Effective Practices for Data. May 20, 2019.
https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp

[5] Schneider, Sarah L. 2018 FDP Faculty Workload Supplement: the University of Michigan. October 2019.

responsibilities as a high priority area for change. Figure 2 below provides more specific information about the data management responsibilities identified as highly in need of change to reduce administrative burden by these 15% of respondents. The blue bar represents the respondents from U-M (Inst. MV1) and the black bar (Categ. MV) represents the six other institutions in the "very high research universities with medical schools" category who also participated in the survey. As shown in the figure, 50% or more of respondents indicated that each data management component listed was a high priority for change. In addition, faculty at U-M listed developing data management plans, institutional resources for data sharing, and deploying information security plans to satisfy applicable laws and regulations as higher priorities for change than did faculty at comparable institutions.

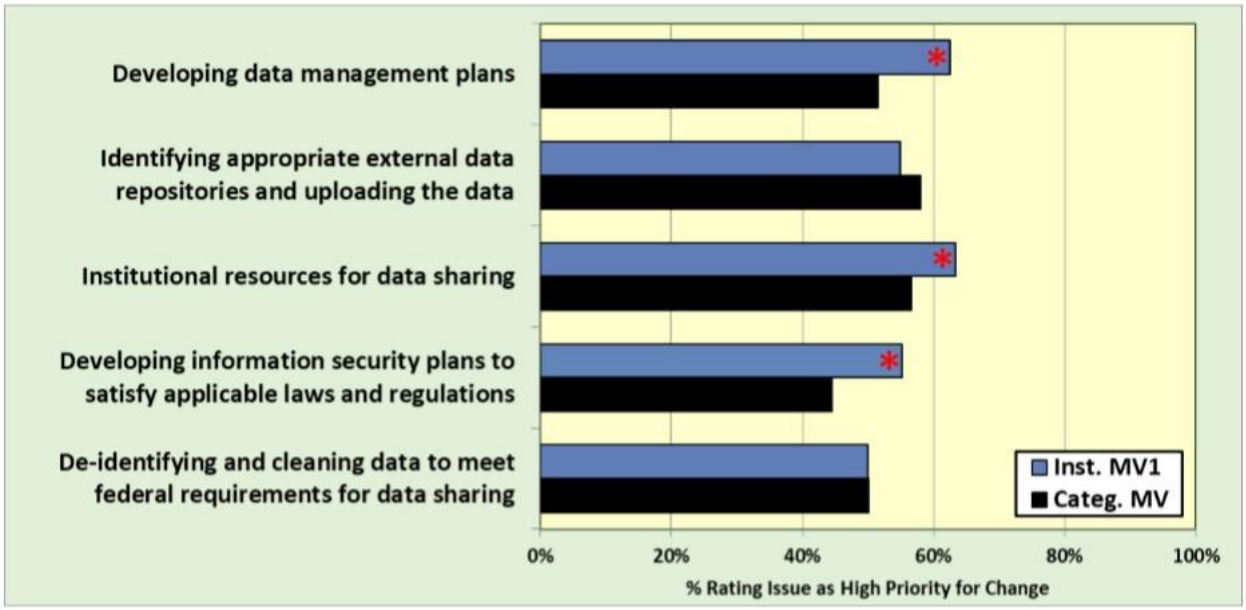


Figure 2 - Percent who rated each data management component as high priority for reducing administrative burden for Category MV and Institution MV1.

Maintaining our preeminence depends upon our ability to continue our success in obtaining funding from federal and private agencies. If U-M does not take steps to ensure that researchers have the guidance, resources, and support they need to comply with the data sharing requirements of funding agencies, we run the risk of losing awards to other research institutions who are able to demonstrate full compliance. We also potentially harm our ability to recruit strong faculty who come to expect support from the University in making their data available to their funding agencies, their research community, and the public in ways that enhance their professional reputation.

**Areas of Focus**

From an institutional standpoint, we believe that there are three broad and interrelated challenges to U-M in providing the support that researchers will need to facilitate making their research data publicly accessible. They are:

- the lack of incentives for researchers to invest time, energy, and resources in sharing their data
- the scale and heterogeneity of data produced by U-M researchers
- the lack of common understanding and connections between units providing research support

*Lack of Incentives for Researchers*

As a premier research institution, U-M has the capacity and responsibility not just to address the challenges of enabling public access to research data, but to lead and serve as a model for our peers. Offering incentives to researchers to invest time, energy, and resources in making their data publicly accessible would reframe data sharing as a worthwhile endeavor rather than a burden. We envision an environment in which high-quality data sets developed by U-M researchers are recognized and rewarded by the institution as first-class products of research, on par with journal articles or books. Creating, communicating, and preserving high-quality datasets requires a significant investment of time, knowledge, and resources on the part of researchers and staff. These investments should be valued and incentivized by the University, particularly in making promotion and tenure decisions.

*Scale and Heterogeneity of Data*

There are a variety of initiatives at U-M that are already working to create a supportive environment for making data more accessible and useable by the public. For example, almost sixty years ago U-M created and continues to host ICPSR[6], the world's largest academic, domain-based data repository. As a result, U-M has been a global leader in the production and sharing of data in the social and behavioral sciences. Data from ICPSR are used by thousands of researchers, policy makers, and others to build knowledge. More recently, the U-M Biological Station developed and launched its own repository[7] to provide wider access to a rich collection of data created by researchers working at that facility. The repository hosts data sets on a variety of topics, including climate change, plant and animal populations, and water flow and quality, all gathered from sites in northern Michigan. Finally, the U-M Library launched its Research Data Services unit in 2016 to provide support for researchers seeking to make their data more widely available. The Library works with researchers to understand their data management, sharing, and preservation needs and then identifies repositories and other resources that could help them meet their needs. The Library also hosts its own data repository, Deep Blue Data,[8] for U-M affiliated researchers to use if needed.

Although we have made strides in making some of the data that underlie our research publicly accessible, the data that are available are only a fraction of the total and variety of data that are produced at U-M. We envision an environment in which U-M actively builds

[6] ICPSR https://www.icpsr.umich.edu/icpsrweb/

[7] U-M Biological Station's Research Gateway http://biostation.lsa.umich.edu/

[8] Deep Blue Data https://deepblue.lib.umich.edu/data

on its leadership in making research data of value more readily accessible and usable by the citizens of Michigan and by people all over the world as a normative part of the research process. This will entail a greater investment on the part of U-M in assisting researchers at U-M to develop and maintain their own data repositories, or other means of supporting data sharing. Perhaps even more importantly however, U-M should recognize that research communities of importance to U-M extend beyond our institutional boundaries. As other institutions provide support to U-M's ICPSR as an important source of high-quality data to social scientists everywhere, we will provide support to other institutions to facilitate access to data in areas of research important to us. We foresee U-M actively engaging in and leading cross-institutional partnerships to support public access to data in critical areas of research at U-M and beyond.

*Lack of Common Understanding and Connections between Units*

As U-M is a large, decentralized organization, responsibilities and activities for supporting research are widely distributed across multiple units. Furthermore, although U-M already has multiple units engaged in overseeing and administering various aspects of our research operations (see Appendix 1), they each have different levels of awareness of issues and expectations around data sharing and what their roles and responsibilities may be. We believe that this lack of shared awareness and connections among units contributes to the frustrations experienced by researchers in meeting their data sharing obligations. In addition, to the best of our knowledge, the requirements for sharing research data are not specifically tracked or accounted for in the systems used by the University to support research activities. It is therefore difficult to know how successful we are, or are not, in meeting our data sharing obligations to funding agencies. This environment also makes it difficult to get a clear understanding of how much data are being generated, what value it might have, and where the University should make investments in sharing, curating, and preserving data of high value.

Sharing data generated at U-M successfully and at scale will require multiple units across campus to have a common understanding of overall expectations, specific responsibilities of each unit, and where connections and hand-offs will need to be made. To reach this state, U-M should create forums and other spaces where representatives of research support units can come together and interact with each other. Initially, these representatives will need to develop a shared understanding of data sharing requirements and how they will work together to provide support to researchers subject to these requirements. Over time, we see units adopting a proactive stance toward data sharing through their experience of working together and anticipating the needs of researchers. Through collaborating with each other, units will develop guides, tools, and other resources to aid researchers in making their data publicly accessible and increase the capabilities of U-M to demonstrate compliance.

We recognize that these challenges will not be solved overnight. It will require a long-term and an ongoing dedication of time, energy, and resources from across the university. Nevertheless, we believe that acting now to develop and promote a culture of data sharing will provide tangible and significant benefits to researchers and to the University over time.

It is in this spirit that we make the following recommendations.

**Recommendation #1 - U-M should create a faculty committee knowledgeable on the challenges of sharing research data and charge them with acting on the issues identified in this report.**

We believe the first step in addressing data sharing issues at U-M would be to form a committee of faculty from across the university with knowledge or expertise on sharing data and to charge them with acting on the issues identified in this report. Though the university already has a large number of committees, charging a committee with understanding and addressing data sharing signifies that U-M sees data sharing as an important issue worthy of everyone's attention.

Although U-M has responded to the data sharing requirements of funding agencies through providing services to researchers, there is no centralized authority on campus to consider the full depth and breadth of challenges that these requirements pose to U-M, and to develop the needed approaches and strategies to address them. The lack of a central, authoritative body has meant that U-M researchers have had to figure out how to respond to data sharing requirements largely on their own. The units on campus that provide services and support for sharing data will need to be a part of the effort to develop a university-wide plan for supporting public access to research data generated at U-M, but faculty will need to lead this effort.

The charge of U-M's Committee on Data Sharing should be written broadly and grant the authority needed to affect university policy, investments and actions. Membership on the committee should include representatives from disciplines and colleges across campus, as well as from UM-Dearborn and UM-Flint. Members of the committee should have some knowledge or expertise in making data accessible beyond their laboratory or office.

We see three initial areas of responsibility to include in the charge for this committee:

- First, this committee should take steps to develop an understanding of researcher experiences across U-M in responding to data sharing requirements.

- Second, this committee should apply what they learn towards developing and promoting a shared set of institutional values and norms at U-M around data sharing.

- Third, this committee should address the gaps and inadequacies of U-M's policies as they pertain to managing, sharing, curating and preserving research data.

*Developing an Understanding of Researcher Needs*

Over the course of this working group we delved into many issues surrounding how the University could support researchers in sharing their data. We had the pleasure of speaking with Dr. Arthur Lupia, U-M's Hal R. Varian Professor of Political Science, who is currently

serving as the Assistant Director of the Social, Behavioral and Economic Sciences (SBE) division of the National Science Foundation. We also spoke with Jason Tallant, the data curator of the U-M BioStation, to get his perspective as someone who works closely with faculty and students in helping them manage, share, and preserve their data. These discussions were critically important in shaping our thinking around faculty and student engagement. However, we recognize that more needs to be done to build a deep and comprehensive understanding of the work life of faculty, staff, and students who have responsibilities for managing, sharing, curating, or preserving research data. Ultimately, the services and support offered by U-M will need to address researchers' day-to-day challenges in order to be effective.

The Committee on Data Sharing would build this understanding by overseeing a deep-dive engagement effort of interviews, observations, or other explorations into understanding the day-to-day experiences of faculty, staff, and students in managing, sharing, curating, and preserving research data. Ideally, the deep dive would include representation from as many of U-M's colleges as possible, though the scale and heterogeneity of the data generated at U-M make it difficult to represent all types of research, disciplinary communities, and data. Therefore, we believe that the focus of the deep dive should be on identifying researchers who have developed systems for managing their data and who are exemplars in making their data publicly accessible. The goal of the deep dive would be to understand researchers' day-to-day data management and sharing practices through seeing first-hand what has worked for them and what barriers remain unsolved. In particular, we seek to learn when and how researchers make use of the services and resources provided by U-M across the lifecycle of their data.

The information gathered from the deep dive would be used to identify themes and patterns to serve as a foundation for an action plan in developing guidance, tools, resources, and services for the larger U-M community. The results would also be used to develop compelling stories to both broaden and deepen discussions on data sharing around campus. Stories, particularly stories from respected peers, are likely to be more illustrative, compelling, and useful in advancing support and services for making data publicly available than top down messages from University administration.

We recognize that a deep dive approach will require a significant investment of time and dedication to achieve meaningful results and that this investment may be more than a University task force can undertake on its own. Therefore, we recommend that U-M consider hiring an outside agency to undertake this work with close coordination and guidance from this task force. ITHAKA S+R[9] is one example of a consulting and research organization with experience in studying data communities that U-M might consider bringing in to run the deep dive.

*Promoting a Shared Set of Institutional Values and Norms*

Researchers look to their field and the scholarly communities to which they belong in establishing norms of practice and for guidance in following these practices. However,

[9] ITHAKA S+R https://sr.ithaka.org

researchers conduct their work at the institutions that employ them and look to their institutions for resources, incentives, and support. Though our focus is local in helping U-M researchers share their data, this work must be informed by expectations from external stakeholders (funding agencies, publishers, etc.) and scholarly disciplines (established best practices, models from exemplars, etc.). Providing efficient and effective support to researchers who are seeking to make their data publicly accessible will require a variety of approaches and actions at different levels of the university. Nevertheless, we believe that U-M would benefit from promulgating a common set of general values to serve as a shared framework for ongoing discussions, investments and actions on sharing data.

The work of the Committee on Data Sharing in developing and promoting specific institutional values will be informed by the results of the deep dive engagement. However, we recommend that U-M consider adopting the following values as a starting point:

1. U-M recognizes that sharing data is necessary for scientific and scholarly practice. It further recognizes that developing and sharing data is a contribution to the advancement of knowledge that merits consideration in tenure and promotion reviews.

2. U-M supports researchers in making their research data accessible. In cases where data cannot be made openly available, consideration should be given to making the data accessible to selected individuals with responsible protections for research purposes.

   There are legitimate legal, privacy, and ethical reasons why data, particularly data on human subjects and other sensitive research areas, cannot be made public. However, there is also a danger of being too conservative or overly cautious in our thinking. Public agencies and foundations supporting research that generates sensitive data are increasingly asking researchers to share their data, as evidenced by a recent notice from the National Institute of Mental Health.[10] This notice includes statements to that effect: "All applications involving human subjects that are submitted to or referred to NIMH are expected to include a Resource Sharing Plan as part of the application" and "Informed consent documents should describe how study data will be shared with NDA (NIMH Data Archive) and the research community."

3. U-M recognizes that sharing data requires a significant investment of time, labor, and resources. The level of investments made by the University will be intentional, guided by University mission and strategy, and informed by anticipated impact.

   The work of developing services and support for data sharing by researchers should be guided by three central questions:

[10] Notice of Data Sharing Policy for the National Institute of Mental Health
https://grants.nih.gov/grants/guide/notice-files/NOT-MH-19-033.html

- Where can U-M best invest resources to enable researchers to share their data in the most efficient, effective, and seamless ways possible?
- How can U-M increase the benefits of data sharing for the researcher and the University as a whole?
- Where can U-M partner with others or adopt external innovations to address common challenges in data sharing?

4. U-M recognizes that sharing research data effectively requires more consideration and investments than just providing access to it once the research is complete. Data sharing requires active decision-making, accounting for costs, and planning *throughout* the life cycle of its development to prepare data for later release with enough contextual information for people to be able to discover, understand, trust and make use of the data. The long-term curation and preservation of research data must also be accounted for through planning.

*Revisiting U-M's Policies on Research Data*

We believe that U-M's policies as expressed through the Standard Practice Guides (SPGs) and other sources do not adequately account for or support managing research data as a distinct type of data requiring support from the university. Currently in U-M's SPGs, research data is considered to be a kind of institutional data. Although this is true at a very broad level, grouping research data with administrative, clinical, and other types of data generated by the University obscures important distinctions and challenges presented by research data. Governance around research data is not as clearly articulated or delineated as it needs to be to support public access to research data effectively. There is ambiguity over ownership rights and responsibilities researchers have over the data they generate. It is often not clear to them what decisions they are expected or empowered to make with their data and what responsibilities the University will need or want to assert.

The charge for the Committee on Data Sharing should include an examination of how U-M policies define and support research data as a distinct category of institutional data. If warranted by this examination, the committee should work to develop a new SPG to address how research data are to be managed, administered, and shared. Current SPGs may not go far enough in distinguishing research data from other types of institutional data. This lack of distinction makes it difficult to address the specific nuances, functions, and issues presented by managing, sharing, and preserving research data.

As we reconsider our policies on research data, efforts should be made to revitalize how data governance is managed across U-M. This includes recognizing that U-M policy and practices on sharing research data may require some flexibility to accommodate and support a variety of disciplinary norms and cultures of practice.

Once research data sharing policy is settled, the Committee should take action to make these policies more visible outside of the SPG context. Researchers need to be made more aware of University policies around research data and the support and services that are available to them in following University policy. Guidance on how policies should or could

be applied to research practice will need to be produced and released in conjunction with any developments to University policies on data. In particular, the Committee on Data Sharing should work to:

a. Clarify roles and responsibilities on ownership and authority over research data. Clearly articulate who has specific decision-making authority over sharing research data and define the larger framework for how decisions are made.

   In general, U-M owns all research data generated or acquired by University faculty, staff, and any other employees (barring any contractual terms to the contrary). It therefore has a responsibility to provide resources and support to ensure research data, as an institutional asset, are properly managed, protected, and preserved.

   Researchers are typically the stewards of the data that they have generated or acquired for research purposes. In these cases, the Principal Investigators (PIs) should have a say in determining how to respond to requirements on sharing research data, including placing research data in public repositories, unless specific terms of sponsorship or other agreements supersede this right.

   The University of Minnesota's (UMN) policy on research data[11] could serve as a model for U-M in clearly articulating expectations and provisions for making data publicly accessible. UMN's policy clearly defines ownership, stewardship, and other responsibilities for research data. It also pulls together relevant policies in other documents and links to them, creating a centralized space where researchers, administrative staff, and others can go for information.

b. Develop provisions to better support the use of U-M's administrative, clinical, or student data for research purposes, bolstered by appropriate regulations, University procedures, ethical considerations, and other safeguards. U-M is already a leader in developing such safeguards through the work of the Institute for Research on Innovation and Science (IRIS)[12] and the Learning Analytics Data Architecture (LARC).[13]

**Recommendation #2 - The units that provide support for data management, sharing, curation, and preservation at U-M should coordinate and align their work even more closely.**

We recommend that U-M create a Research Data Sharing Service Group (RDSSG) comprised of representatives from units that provide support for data management and sharing activities from across U-M. There are many units at U-M that support its research

[11] University of Minnesota's policy on research data https://policy.umn.edu/research/researchdata

[12] IRIS https://iris.isr.umich.edu/

[13] LARC https://enrollment.umich.edu/data-research/learning-analytics-data-architecture-larc

operations, including some that focus on assisting researchers in managing, sharing, curating and preserving their data (see Appendix 1). However, no single unit can provide the full range of support and services across the data life cycle that are needed for researchers to share their data effectively. The RDSSG would promote closer connections among these units and enable more of a "one university" response to the challenges presented in sharing data.

The RDSSG would serve as a means for units to:

1. communicate on areas of mutual interest through cross-campus discussions
2. coordinate on initiatives to market and raise awareness of new or existing services
3. heighten the visibility of support available for data by providing a centralized presence to researchers and others at U-M
4. continue to identify areas of need at U-M
5. create a stronger voice at U-M in communicating which service gaps should be addressed and where investments would be most beneficial
6. work closely with the faculty Committee on Data Sharing to assist them in carrying out their responsibilities as appropriate.

The charge of the RDSSG should include leading the following initiatives. We recommend that the RDSSG work closely with the faculty Committee on Data Sharing in exploring and carrying out these initiatives.

*Inventory of Services*

The RDSSG, working with as many of the units that provide support for data management and sharing activities as possible, should develop an inventory of the services offered to researchers. This will create a shared understanding of the full range of services, raising awareness among service providers. Developing an inventory of services will also better enable referrals between agencies and promote a shared sense of community. The MI Research Cores website[14] provides an example of how a centralized portal for connecting researchers to services could be done. However, this website was only recently introduced, and it is not yet clear how successful it has been. Outreach efforts will need to include metrics for success, informed by the needs of researchers (see recommendation #1), and the means to get feedback from users and non-users alike.

*Better inventorying of U-M's Research Data*

The RDSSG should launch an exploration of how U-M could better account for and inventory the types and amounts of research data generated by U-M researchers, including our understanding of where the data resides, and who has authority and stewardship responsibilities over which data. This would enable the University to better determine where to prioritize investments and identify where additional support may be needed.

[14] MI Research Cores website http://cores.research.umich.edu

However, we should recognize that the University is comprised of a wide and diverse set of fields and that the dynamism of data sources means that there is a limit to our ability to catalog and account for all data across the University. The departments and centers of U-M will need to build their capacity incrementally and in ways that address the needs of faculty and departments (and not just University administration). This effort will likely be more effective if it is initiated at smaller scales at the level of schools or colleges with the backing of University administration.

*Developing Actionable Data Management Plans*

The RDSSG should explore the potential of Machine Actionable DMPs (maDMPs), in which the DMP is incorporated into the tracking and reporting systems U-M uses to administer a grant. DMPs are often developed in a hurry and without much consideration for how they will be implemented. After the grant has been submitted, DMPs are often forgotten until the award period ends and the data must be made available. MaDMPs are being promoted by the data curation community[15] and by funding agencies[16] to overcome the shortcomings of stand-alone DMPs. The RDSSG should learn about the benefits and costs of maDMPs and provide a recommendation to U-M administration as to whether maDMPs should be adopted by the university.

The Michigan Institute for Clinical and Health Research (MICHR) already carries DMPs forward as living documents in the research life cycle to inform compliance, reporting, and publishing. Their process may help to inform how we might make DMPs more useful and actionable by researchers and administrators across U-M.

*Setting up and making use of digital storage space*

The RDSSG should consider the need for accessible and affordable high-quality storage systems for sharing and preserving the data generated at U-M and how these needs may grow or shift over time as data sharing requirements evolve. We have observed that researchers will often choose the more cost-efficient options for storing their data, particularly options offered by commercial retailers, rather than higher-quality options offered by the university. This places data at risk and introduces additional complexities transitioning data into environments where they will be shared, curated, and preserved. Incentives to researchers for actively managing their data -- with sharing and reuse in mind -- by periodically cleaning it and ensuring its fitness for use with new software tools should also be considered.

The RDSSG should also consider a means of provisioning the costs of long-term storage systems for units, such as the library, who host University data sets for sharing, curation, and preservation.

---

[15] Miksa T, Simms S, Mietchen D, Jones S. "Ten principles for machine-actionable data management plans." PLOS Computational Biology 15(3): e1006750 (2019). https://doi.org/10.1371/journal.pcbi.1006750.

[16] Dear Colleague Letter: Effective Practices for Data. May 20, 2019. https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp.

Educational Programming

Working closely with the faculty Committee on Data Sharing, the RDSSG should foster and promote educational programming for U-M faculty, staff and students. Increasing public access to research data depends upon educating those who are responsible for generating, managing, or stewarding data sets so that they know how to manage and share data and understand the benefits of doing so. A recent study of social science graduate curricula across 140 programs found almost none included training on data management or data sharing.[17]

U-M is a complex and diverse institution. As such, it will require a multitude of educational programs to reach and connect with faculty and students. Educational programming should be designed to reach specific audiences and include disciplinary norms, standards, and practices to the extent that is possible. The RDSSG should support the development of education and training materials that are effective for its many different communities and encourage sharing of those resources across the University and beyond.

There are some educational programs across U-M that already incorporate in their curricula issues related to sharing research data. The Responsible Conduct of Research training program is offered by the University and required of those seeking funding from the National Institutes of Health, the National Science Foundation, and other funding agencies. As described on U-M's RCR training website, successful training programs include instruction in "data acquisition and laboratory tools; management, sharing and ownership."[18] The University also has a strong Software and Data Carpentry[19] community led by Patrick Schloss, professor of Microbiology and Immunology. Both of the carpentries teach essential computational skills sets to students seeking to do data-driven research and include data management as an area of focus. Lesson plans for these programs could be reviewed and augmented as necessary to include specific information or resources to prepare researchers to make their data available.

It would be useful to have a better understanding of how many people are reached through existing educational programming and how these programs have influenced research practices. New educational programing could be created to reach audiences who are not currently served by existing programming or to address unmet needs. One possible program would be a workshop for incoming faculty on considerations for setting up their new research lab.

---

[17] Ashley Doonan, Dharma Akmon, and Evan Cosby (2019) "An Exploratory Analysis of Social Science Graduate Education in Data Management and Data Sharing" available at: http://hdl.handle.net/2027.42/150174.

[18] Responsible Conduct of Research (RCR) Training https://research-compliance.umich.edu/responsible-conduct-research-rcr-training

[19] The Carpentries https://carpentries.org/

*Support for Assigning Persistent Identifiers*

The RDSSG should support the adoption of the use of persistent identifiers in research management systems, where it is possible and appropriate to do so. The RDSSG should also promote the importance of using persistent identifiers to the U-M community. Persistent identifiers are becoming increasingly essential in enabling scholarly communication. Unique identifiers assigned to a person (such as an ORCID[20]), a work (such as a DOI[21]), or an institution (such as a ROR[22]) facilitate the findability of journal articles, data sets, or other products of research across distributed networks. Using identifiers also enables tracking research outputs. This makes it easier to comply with data sharing requirements and to support proper attribution to the researchers and the institutions where the work was created.

*Publishing data / Depositing data into a repository*

The RDSSG should assist U-M in adopting tools to facilitate the deposit of articles and data by researchers into repositories that are compliant with funding agency requirements. Compliance tools, such as the Public Access Submission Service (PASS)[23] from Johns Hopkins University or Chronos,[24] are being developed to make the deposit of articles into compliant repositories easier for researchers to do. These tools are further expanding their efforts into data. Where appropriate, e should collaborate with agencies who share our values to help them extend and strengthen their tools' capabilities and then implement these tools at U-M.

[20] Open Researcher Contributor IDentifer (ORCID) http://orcid.org

[21] Digital Object Identifier (DOI) http://doi.org

[22] Research Organization Registry (ROR) https://www.ror.community/

[23] PASS http://pass.jhu.edu

[24] Chronos http://chronos-oa.com

**Appendix 1: Existing Support Services at the University of Michigan**
There already are multiple units at U-M providing critical services to researchers in managing, sharing, curating, or preserving their data in some form, including:

- **Advanced Research Computing - Technology Services (ARC-TS)** offers access to and support for a variety of computing infrastructure, including cloud computing services, research storage, and high-performance computing systems.
- **Consulting for Statistics, Computing, and Analytics Research (CSCAR)** provides support and training for the U-M community on the management, collection, and analysis of data and manages the Data Acquisition for Data Science (DADS) initiative with the Library to provide seed funding for broadly used specialized datasets
- **The Data Office for Clinical and Translational Research (DOCTR)** supports medical research by enabling secure access to patient data through a variety of tools or delivering customized data sets based on researcher needs.
- **ICPSR** hosts a member-based data archive of more than 500,000 files, which includes specialized collections of data in education, aging, criminal justice, substance abuse, and other fields. ICPSR hosts many educational courses, notably its summer program in Quantitative Methods of Social Research, and supports the use of data in teaching students. ICPSR also conducts research on the evolving challenges of sharing, curating, and preserving research data and how data hosted by repositories are used by others.
- **Information Technology Services (ITS)** is charged with securing, protecting, and backing up the data generated and collected by U-M.
- **The U-M Library** offers services to support U-M researchers seeking to manage, share, curate, or preserve their research data. These services include reviewing researchers' data management plans, educating on best practices, and working with researchers to help them deposit their data into publicly accessible repositories. In 2016, the library launched Deep Blue Data, an institutional data repository for U-M researchers whose research community had not yet developed a data repository of their own. The library is a founding member of the Data Curation Network, a collaboration of ten institutions to share data curation staffing and expertise to more effectively curate a wide variety of data formats and types.
- **The Michigan Institute for Data Science (MIDAS)** serves as U-M's primary vehicle for developing and supporting data science initiatives. MIDAS stewards a number of research data sets for use by the U-M community and is developing an online platform for ongoing collaboration and engagement.
- **The Office of Research and Sponsored Programs (ORSP)** oversees the conduct of research and sponsored activity at U-M throughout the life cycle of the project. The office is responsible for ensuring that researchers are complying with funding requirements around managing, sharing, and preserving data.
- **The Research Ethics and Compliance Office** provides information, guidance, and oversight functions regarding federal regulations and funding sponsor requirements, including the protection of human and animal subjects used in research, export controls, and protections for controlled unclassified information. This office also has responsibility for overseeing and enforcing U-M's Research Integrity program and responsible conduct of research (RCR) training.

## Appendix 2: Initiatives and Programs of Peer Institutions

- **Cornell University:** Soon after the NSF's announcement about their data management plan requirement in 2011, Cornell launched its Research Data Management Service Group (RDMSG) comprised of staff from the Library, IT and computing units, and a social science data archive. According to their website, the RDMSG "is a collaborative, campus-wide organization that assists with creating and implementing data management plans, applying best practices for managing data, and finding data management services at any stage of the research process." In addition, the RDMSG develops tools to address the data management and sharing needs of Cornell researchers. Most recently, they developed the "Data Storage Finder" tool to assist researchers in identifying which of Cornell's storage options is right for their data. https://data.research.cornell.edu/

- **Massachusetts Institute of Technology (MIT):** MIT's Open Access Task Force recently released its report in which it advocates for increasing support for open sharing of research output, including data. https://open-access.mit.edu/sites/default/files/OA-Final-Report.pdf. In 2015, the MIT libraries convened a faculty task force and charged them with "developing a vision of how the MIT Libraries ought to evolve to best advance the creation, dissemination, and preservation of knowledge." Their report included discussions on the library as a steward for the research outputs of MIT, including data. https://future-of-libraries.mit.edu/sites/default/files/FutureLibraries-PrelimReport-Final.pdf

- **Purdue University:** Purdue University Libraries began their program of applying library science to data management in 2005. They created their Distributed Data Curation Center (D2C2) as a means to foster collaboration and research in data curation. The Purdue University Research Repository (PURR) not only serves as a repository for sharing and preserving data sets but provides researchers with collaboration tools and space to develop and manage their data prior to sharing it. https://www.lib.purdue.edu/researchdata

- **Stanford University:** After the AAU-APLU meeting in Oct 2018, Stanford assembled a cross-campus Public Access to Research Data (PARD) working group that includes representatives from the library, medical school, dean of research, and research computing. The group is informally charged to develop recommendations to the Vice Provost and Dean of Research on strategies for ensuring Stanford plays a leadership role in the issues around data access and stewardship.

- **The University of California - Berkeley:** UC-Berkeley will soon be unveiling its campus data portal to centralize the services provided to researchers to manage and share their data. https://test-bigr-data.pantheon.berkeley.edu/home-alt

- **University of Minnesota:** The University of Minnesota's policy on research data is a strong model for clearly articulating expectations and the support available for

managing and sharing data. https://policy.umn.edu/research/researchdata. The UMN is also the lead institution of the Data Curation Network (DCN). The DCN is a consortium of ten institutions (including U-M) who share the expertise of their data curators across institutional boundaries to enable its members to take advantage of a broader pool of curation expertise than any one institution could provide on its own. http://datacurationnetwork.org