Realities of Academic Data Sharing (RADS) Initiative:

Research Methodology 2022–2023 Surveys and Interviews

January 2024

https://doi.org/10.29242/report.radsmethodology2023

ASSOCIATION OF RESEARCH LIBRARIES

U.S. National Science Foundation

Authors

Shawna Taylor
Association of Research Libraries
ORCID: 0000-0002-9842-7867

Joel Herndon
Duke University
ORCID: 0000-0001-9995-9040

Alicia Hofelich Mohr
University of Minnesota
ORCID: 0000-0002-7644-4105

Wendy Kozlowski
Cornell University Library
ORCID: 0000-0001-6539-3798

Jonathan Petters
Virginia Tech
ORCID: 0000-0002-0853-5814

Jennifer Moore
Washington University in St. Louis
ORCID: 0000-0001-6628-6820

Jake Carlson
University at Buffalo
ORCID: 0000-0003-2733-0969

Cynthia Hudson Vitale
Association of Research Libraries
ORCID: 0000-0001-5581-5678

Lizhao Ge
George Washington University
ORCID: 0009-0005-7862-6016

This material is based upon work supported by the US National Science Foundation grant number 2135874.

Suggested citation: Taylor, Shawna, Alicia Hofelich Mohr, Jonathan Petters, Jake Carslon, Lizhao Ge, Joel Herndon, Wendy Kozlowski, Jennifer Moore, and Cynthia Hudson Vitale. *Realities of Academic Data Sharing (RADS) Initiative: Research Methodology 2022–2023 Surveys and Interviews*. Washington, DC: Association of Research Libraries, January 2024.

https://doi.org/10.29242/report.radsmethodology2023.

Table of Contents

Overview

This report describes the methodology of research conducted during the first stage of the Realities of Academic Data Sharing (RADS) Initiative[1], funded by the US National Science Foundation (NSF), from 2021 to 2023, and should be considered supplemental to the additional final research reports (white papers) produced as a result of this research. As part of the RADS Initiative, institutional administrators and funded researchers were surveyed in 2022 and interviewed in 2023 on details related to research data sharing support services and practices, and their corresponding expenses. While the Association of Research Libraries (ARL) is the administrative home of the RADS Initiative, the research was conducted with participants at the following institutions: Cornell University, Duke University, University of Michigan, University of Minnesota, Virginia Tech, and Washington University in St. Louis.

1 The first stage of the RADS Initiative was funded by the US National Science Foundation (NSF), award number 2135874, and the second stage of the initiative has been funded by the US Institute of Museum and Library Services (IMLS), award number LG-254930-OLS-23. All research described in this report is an output of first-stage NSF funding.

Introduction & Research Purpose

Increasing federal requirements for funded researchers to share their research data for public access has increased over the last 10 to 15 years, particularly since the release of the 2013 White House Office of Science and Technology Policy (OSTP) Holdren Memo "Increasing Access to the Results of Federally Funded Scientific Research."[2] As requirements have increased, many academic institutions have developed and launched a variety of support services to reduce their researcher burden in meeting these requirements. Services are often dispersed across the institution and, as a result, the extent of these services and the costs to the institution to support these services, have not yet been understood. Similarly, the extent to which funded researchers undertake activities to make their data publicly accessible, and their expenses for data sharing, has not yet been explored. The goal of the first stage of research in the RADS Initiative, conducted from 2021 to 2023, was to better understand these activities and costs to institutions and funded researchers.

The methodology described in this paper pertains to two research questions[3] considered during the first research stage of the RADS Initiative:

1. How are researchers making decisions about why and how to share research data?
2. What is the cost to the institution to implement the federally mandated public access to research data policy?

---

2 John P. Holdren, "Increasing Access to the Results of Federally Funded Scientific Research," Office of Science and Technology Policy, Executive Office of the President, February 22, 2013, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

3 A third research question, "Where are funded researchers across these institutions making their data publicly accessible and what is the quality of the metadata?" was also included in the initial funding stage of the RADS Initiative. This stream of inquiry is outside the scope of the methodology described in this report. For methodology and analysis of this research question, see the forthcoming published article by the RADS team, which will be listed on the Realities of Academic Data Sharing (RADS) Initiative website, https://www.arl.org/realities-of-academic-data-sharing-rads-initiative/.

The RADS study was retrospective, investigating data sharing and support activities from 2013 to 2022, and consisted of surveying and interviewing institutional administrators with expenditure knowledge of their departments/units and funded researchers at the six participating institutions. Project principal investigators (PIs) at their respective institutions submitted applications to their institutional review boards (IRBs) for study approval. IRBs at each institution approved the study or deemed it not applicable under the human participants category. The following members of the research team were the PIs at their respective institutions and the institutional IRB points of contact.

- Jake Carlson, formerly the director of Deep Blue Repository and Research Data Services, University of Michigan (currently the associate university librarian for Research, Collections & Outreach, University at Buffalo Libraries, University at Buffalo, as of August 2023)
- Joel Herndon, director of the Center for Data and Visualization Sciences, University Libraries, Duke University
- Alicia Hofelich Mohr, Research Support Services coordinator, Liberal Arts Technologies and Innovation Services (LATIS), University of Minnesota
- Wendy Kozlowski, director, Research Data and Open Scholarship, Cornell University Library, Cornell University
- Jennifer Moore, head of Data Services, University Libraries, Washington University in St. Louis
- Jonathan Petters, assistant director, Data Management & Curation Services, Data Services, University Libraries, Virginia Tech

Goals of the research included institutional scans of data management and sharing activities, collecting information on data sharing activity expenses for researchers and administrators, and an assessment on the impact of data management and sharing policies to both the researcher and institution based on data from qualitative interviews.

Terminology

Data Sharing

Within this report "data sharing" practices, broadly speaking, may include researchers sharing data upon request, limited access or restricted sharing, or sharing on platforms without restrictions and available to anyone. Although the RADS study is interested in data sharing for public access, the questions in the surveys inquired into the broad sharing of federally funded research data. Defining data sharing in the surveys in the broadest sense was intentional, as data sharing likely means different things across disciplines and institutional roles, from placing data in public repositories to providing access to restricted storage.

Infrastructure

The term "infrastructure" in this report, and throughout all of our research outputs, is used as a singular term to encompass all institutional efforts to support research data sharing and management activities, broadly speaking. This includes: technical infrastructure (such as institutional repository support); data governance, including the development, implementation, and oversight of data policies; one-time efforts or investments to accelerate services; and ongoing service operations. Staffing time and costs, while essential to implement and maintain infrastructure and run services, are considered as a separate category in our analysis

Survey Development: RADS Public Access to Data Sharing and Management Activities

In order to inquire into data sharing practices on the surveys, the RADS survey team developed a list of activities that would serve as a common grounding, or shared vocabulary, of the concrete actions involved in managing and sharing data (with public access data sharing specifically in mind).[4] The study team devised two lists of activities, which included 28 activities for the researcher audience and 27 activities for the administrator/research support audience. These activities were placed into five data sharing life-cycle phases, influenced by the research and grant life cycles. The full list of activities for both participants are listed in Appendix A of this report, as well as in the ARL report, Public Access Data Management and Sharing Activities for Academic Administration and Researchers (November 2022).

These activities were developed in collaboration with COGR, who, at the time, was developing a Roles & Responsibilities list for their NIH Data Management and Sharing Readiness Guide.[5] Other frameworks we consulted when developing the data sharing activities include:

- "Cost-Benefit Studies, Tools, and Methodologies Focusing on Long-Lived Data," Keeping Research Data Safe (KRDS), https://beagrie.com/krds.
- Data Management Costing Tool and Checklist, UK Data Service, 2022, https://dam.ukdataservice.ac.uk/media/622368/costingtool.pdf.

4 Some data sharing activities for public access, such as sharing on an open platform, can be assumed to always be required to enable data sharing; however, there are many activities that are only utilized depending on the attributes of the data or data type (for example, large datasets may require additional considerations for transfer, or data sharing may be restricted due to reuse agreements, etc.). The RADS data sharing activities lists considered the widest range of activities related to data sharing.

5 See "Chapter 3—Implementation Roles and Responsibilities, Part II, Roles & Responsibilities," in COGR's NIH Data Management and Sharing Readiness Guide, COGR, November 8, 2022, https://www.cogr.edu/cogr-readiness-guide-chapter-3-implementation-roles-and-responsibilities.

- "Chapter 2, Framework Foundation: Data States and Associated Activities," Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs, A Consensus Study Report of The National Academies of Sciences, Engineering, and Medicine (Washington, DC: The National Academies Press, 2020), https://nap.nationalacademies.org/read/25639/chapter/4.
- Total Cost of Stewardship: Responsible Collection Building in Archives and Special Collections (Dublin, Ohio: OCLC Research, 2021), https://doi.org/10.25333/zbh0-a044.

These activities were used to gauge both the extent to which researchers and administrators were engaging in or supporting these actions, as well as the associated cost in terms of staffing and technical infrastructure.

Survey Methodology

Institutional Infrastructure Survey for Administrators

Administrator Participant Pool Identification

To determine which departments/units (hereinafter referred to as "units") to survey, and who in particular to survey, each RADS PI conducted a scan of their institution to identify possible units that support funded researchers with any data sharing activities (as identified in our survey development process). In addition to leveraging their personal institutional knowledge to begin the scan, PIs contacted known administrators whose units supported data sharing to inquire into other possible offices to include in the survey pool, and also used institutional websites to identify units to include in the survey pool.

After this scan, administrators of these units were identified and then invited to participate in the survey. Additional participation inclusion criteria included:

- Knowledge of department/unit infrastructure expenditures
- Knowledge of personnel activities to support data sharing
- Knowledge of personnel salaries

The number of identified offices/administrators in the participant pool varied among the six RADS institutions, from 15 to 34 (Table 1). When administrators from multiple units under one department participated, administrators were asked to report on activities and expenditures for their unit only. No restrictions were placed on collaborative unit efforts to complete the survey, and it is known that up to four individuals from one unit worked together to complete one survey response for their unit. On these occasions, only one administrator name was recorded in the survey.

Pilot Institutional Infrastructure Survey

Surveys were sent to 10 pilot participants across all institutions for feedback on the questions, descriptions, and clarity of the survey in August 2022. Changes were integrated into the survey during September 2022 before release to the larger participant pool. Pilot participants were invited to complete the final version of the survey and only their responses to the final version were included in the analysis.

Institutional Infrastructure Survey Release Details

The Institutional Infrastructure Survey (see Research Instrument #1 below) was open from October 3, 2022, to December 5, 2022, and was hosted on the Alchemer platform. All potential participants were emailed a week before the survey opening by the RADS PI of their respective institution to let them know the survey would be coming, and to send a copy of the questions so they could prepare. Survey links were sent individually by each RADS PI; subsequent email reminders (up to three) were sent to administrators who had not responded throughout the open survey period.

Institutional Infrastructure Survey (Administrator) Response Rate

Before analysis, duplicated individual responses were removed when the same individual submitted the survey more than once. When this occurred, the most complete survey response was retained. Additionally, responses from administrators within the same unit were collapsed into a single response. When there were responses from multiple people in the same unit, the most complete survey response was counted. When there were discrepancies in the responses from multiple people in the same unit, we took the response from the most senior respondent. Text from open-ended responses were combined with the retained response, where applicable. Furthermore, participants of the pilot survey were invited to retake the final survey

and, if they did participate, their response was counted in the overall response rate.

After removing and collapsing duplicate responses, the response rate of the administrators ranged from 29.5% to 70.6% across institutions, with an overall average response rate of 50.0% (see Table 1).

Table 1: Response rate of administrators invited to complete the RADS Institutional Infrastructure Survey. Note: The response rates reflected in Table 1 are based on the cleaned data, which does not include demographic-only responses (for example, institution, email).

RADS Institution	Number of Invited Administrators	Number of Responses	Response Rate
Cornell University	17	12	70.6%
Duke University	15	9	60.0%
University of Michigan	22	14	63.6%
University of Minnesota	33	19	57.6%
Virginia Tech	17	7	41.2%
Washington University in St. Luis	34	8	29.5%
Total & Average Response Rate	**138**	**69**	**50.0%**

Unit Categorization

Recognizing institutions vary in their administrative structure and organization, responding offices were categorized into one of four service-based areas to enable comparisons across the six different institutions. These areas are: Libraries (LIB), Central Administrative Research Offices (RSCH), Information Technology (IT), and discipline-specific Institutions or Research Centers (IC). See Appendix B for a list of all responding offices and their service-based category.

Expense Data Cleaning

Of the 69 responses, 58 administrators (84%) provided information about their expenses. Expense data was cleaned to ensure consistency in how responses were entered (such as whole dollar amounts, removing text, and putting in a numeric format). When ranges were given, the median was taken. Individual responses were examined to ensure responses were consistent across questions (for example, ensuring the number of staff reported and the number of salaries/time reported matched).

Where responses were unclear (for example, without annual or hourly demarcations), conflicting (such as reporting salaries but zero employees), contained only partial information, or reported very high expenditures (either in salary or infrastructure costs), administrators were recontacted for clarification via email or in follow-up interviews.

Percentage time and number of staff reported were combined into full-time equivalent (FTE) values by summing up the percent effort of each reported staff member dedicated to data management and sharing. Salary and percent effort were also calculated to provide a total annual cost associated with personnel time supporting data management and sharing activities. Total infrastructure costs were adjusted to remove staffing expenses, as we found in interviews and follow-up queries that participants frequently included both staffing and infrastructure costs in this survey entry. In cases where infrastructure cost exceeded staffing, the staffing number was subtracted out. This adjustment undoubtedly resulted in underreporting, as some respondents may have truly had infrastructure costs that exceeded staffing. Adjusted staffing and recurring infrastructure costs were summed to create a total annual cost.

Institutional Infrastructure Survey Limitations

The following are limitations of the Institutional Infrastructure Survey:

- The population pool of institutional administrators was difficult to create, and RADS project PIs individually identified units thought to support data sharing activities at their respective institutions. PIs may have missed identifying units in their institutional assessments of service providers and, therefore, these absent units would not be represented in our data.
- When developing the administrator participant pool, understanding at what level to survey administrative offices was occasionally challenging (for example, which units had their own budgets within the Office of the Vice Provost for Research), and unit levels varied between institutions.
- Known units/departments that support data sharing did not respond to the survey. Therefore, we know our data, including the resulting interactive visualizations, are incomplete.
- In an effort to reduce survey burden, administrators were first asked about the general phases their unit/department supports or offers services for in question 6; phases were listed with sample activities as examples. Respondents were then only asked about the full list of activities from that phase if they indicated support for it (questions 7–11). Units/departments may support activities that were not shown to them based on the questionnaire structure. This is a possible area of underreporting in our data.
- Questions 7–11 asked administrators about the data management and sharing activities their unit supports, and all responses indicated supporting at least one activity. However, some responses indicated "zero" to question 12, "Approximately how many staff did you hire or rely on to support public access to research data?" We can assume at least one person would be required to fulfill reported data management and sharing activities, even if the number of activities is minimal. When "zero" responses were indicated for question 12, the subsequent question

regarding percent of time dedicated to the activities and salaries was a nonresponse, causing our data to be underreported in these areas.

- During follow-up interviews with a sample size of administrators, we recognized a pattern for question 12, "Approximately how many staff did you hire or rely on to support public access to research data?"—administrators consistently did not include themselves or other senior administrators in their units. Using a broader term such as "personnel" may be more inclusive, or simply asking respondents to include themselves in their responses. Therefore, data for question 13, regarding percent of time toward the activities and salaries, is underreported in these areas.
- For the salary-centered questions (13 and 14), some respondents included benefits/fringe, while others did not.
- There is likely underreporting in our data for question 15, "For services, infrastructure, or staffing costs related to data sharing across the research life cycle, what was your approximate budget (e.g., software, contracts, fees) for 2021–2022? Recurring annual expenses ($)? One time expenses ($)?" Due to the way in which this question was worded, some respondents would have included total staffing costs in their "recurring" response. When the recurring response was higher than the total reported salaries, the project team decided to subtract the salary figure from recurring expenses, so as to not overreport expenses in this area.
- Analyzing annual costs across categories of institutional offices was challenging, as many universities have multiple offices in each of these categories, with independent budgets and costs. Reported expenditures in these broad categories (LIB, IT, RSCH, IC) do not account for the number of offices within each institution.

Researcher Survey

Researcher Inclusion Criteria

To make meaningful comparisons across discipline areas, we restricted the participant pool to five grant/project disciplinary areas: environmental science, materials science, psychology, biomedical sciences, and physics. These specific disciplinary areas were selected to ensure a variety of data sharing practices, across time, would be captured in participant data. The chosen areas were also based on disciplines in which datasets have been submitted to the Data Curation Network (DCN),[6] as well as areas in which each of the participating institutions have a strong research presence.

After the discipline areas were determined, funders were narrowed down to include only the US Department of Energy (DOE), the US National Institutes of Health (NIH), and the US National Science Foundation (NSF). These three funders were selected specifically, as it was determined the majority of federal funding at the RADS institutions came from these agencies and award information was publicly available via their award databases. All inclusion/exclusion criteria for all researchers in the participant pool were:

- Researchers must have been externally funded and their project listed in one of three funder award databases (DOE, NIH, or NSF).
- Researchers must be currently employed at the same institution as when the award was granted at the time of survey.
- Awarded projects must fall into one of the five disciplinary areas of consideration in this study: environmental science, materials science, psychology, biomedical sciences, physics, or a cross-disciplinary study including one of these five areas. Multidisciplinary projects were considered, as long as one of the disciplines included one of these five discipline areas.

6 The members of the Data Curation Network are part of a shared staffing model where, if there is no expertise or capacity at one institution to curate a particular type of dataset, datasets can be submitted to the entire DCN for curation. DCN data from 2019 to 2021 informed the selection of discipline areas. Current DCN data on "Datasets Submitted to the DCN by Discipline" (https://datacurationnetwork.org/data-visualization/) reflect current shared curation data.

- Projects must have been completed between 2013 and 2022, excluding no-cost grant extensions. The year 2013 was selected as the starting point for analysis due to the release of the OSTP-issued Holdren memo.

Research Participant Population

The researcher survey subject population was identified by querying information in the publicly available DOE, NIH, and NSF award databases. Funded researchers from the six RADS institutions were identified from the API pulls, as described below, and possible participants for the survey pool were de-duplicated; when multiple grants were awarded to the same PI, only the latest award information was considered.

National Science Foundation (NSF) Pulls

API calls were made using V1 of the NSF Award Search API.[7] Data were pulled between March 26 and March 29, 2022 (see the RADS NSF R scripts[8]). Institutions were searched by name using the awardeeName parameter. City was used as an additional parameter for Washington University and Virginia Tech in order to narrow down the search to the relevant institutions. Calls were made by page in order up to 10,000 results. No institution had more than 5,000 results. An additional call was made to gather the program name, abstract, and PI contact details for each ID. Awards were then evaluated for inclusion within each of the disciplines using the fund program area. NSF does not have a field to indicate completed awards.

7 "NSF Awards," Research.gov, US National Science Foundation, accessed December 21, 2023, https://resources.research.gov/common/webapi/awardapisearch-v1.htm.

8 Realities of Academic Data Sharing (RADS)—API Scripts, accessed December 21, 2023, https://github.com/DataCurationNetwork/rads-api-pulls/.

National Institutes of Health (NIH) Pulls

API calls were made to the NIH RePORTER API[9] using the repoRter.nih package for R[10] and were pulled between March 29 and March 30, 2022 (see the RADS NIH R scripts[11]). Completed awards from institutions were searched by organizational name between 2013 and 2022. Data from Washington University in St. Louis was additionally narrowed down using organization city in the call. The complete data was pulled for each, with the exception of the University of Michigan and Washington University in St. Louis, which hit the API maximum return of 10,000 awards. Awards were evaluated for discipline using the organizational department type.

Department of Energy (DOE) Pulls

Data was pulled from the DOE Award Search database[12] using the web search interface on April 21, 2022, and results were downloaded into a Microsoft Excel file. Each university was searched by institution name, and subsetted to include inactive grant awards. Data from Virginia Tech and Washington University in St. Louis were additionally narrowed down by city and state, respectively. Grants were included with start dates of 2013 or later. Individual awards were evaluated for discipline using the program area.

Only grants with start dates after January 01, 2013, and end dates before May 01, 2022, were included. For researchers who had multiple grants from these agencies, only the most recent was taken. NIH emails were gathered from publicly available FOIA requested datasets, and then filled in manually using the NIH Reporter website. DOE emails were found through online search. The NSF API provided emails for

9 "NIH RePORTER," US National Institutes of Health, accessed December 21, 2023, https://api.reporter.nih.gov/.

10 Michael Barr, "repoRter.nih: R Interface to the 'NIH RePORTER Project' API," R package version 0.1.1, 2022, https://CRAN.R-project.org/package=repoRter.nih.

11 Realities of Academic Data Sharing (RADS)—API Scripts, accessed December 21, 2023, https://github.com/DataCurationNetwork/rads-api-pulls/.

12 "Award Search," Portfolio Analysis and Management System, Office of Science, US Department of Energy, accessed December 21, 2023, https://pamspublic.science.energy.gov/WebPAMSExternal/interface/awards/AwardSearchExternal.aspx.

grant PIs. Discipline was coded for each potential participant using the PI's department (NIH) or based on the grant title, directorate, and/or award abstract (NSF and DOE). Awards were subsetted to include only awards in the relevant disciplines.

Pilot Researcher Survey

Surveys were sent to eight pilot participants across all institutions for feedback on the questions, descriptions, and clarity of the survey in August 2022. Changes were integrated into the survey during September 2022 before release to the larger participant pool. Pilot participants were invited to complete the final version of the survey and only their responses to the final version were included in the analysis.

Researcher Survey Release Details

The researcher survey was open from October 3 to December 5, 2022, and the Alchemer platform was used to host the survey. All identified researchers from the API pulls were emailed the survey through Alchemer. Each researcher had an identification number (ID) auto-generated through Alchemer; this ID number allowed the study team to send two follow-up emails through Alchemer to nonresponding individuals in the participant pool.

Of an initial 263 survey responses, 8 were removed due to the following factors:

- The response was not considered complete, as only demographic questions (such as email, department) were answered.
- The response was a duplicate response by the same participant.
- The respondent was no longer affiliated with the institution (retired or otherwise left the institution).

Researcher Survey Response Rate

After the data cleaning described above, this brought the total number of responses to 255. The overall response rate of the researchers varied by institution, from 4.9% to 14.0% (see Table 2 below), with an average response rate of 8.4%.

Table 2: Response rate of researchers invited to complete the RADS Researcher Survey. Note: The response rates reflected in Table 2 are based on the cleaned data, which exclude retirees and researchers no longer affiliated with the RADS institutions.

RADS Institution	Number of Invited Researchers	Number of Responses	Response Rate
Cornell University	312	28	9.0%
Duke University	618	40	6.5%
University of Michigan	1003	49	4.9%
University of Minnesota	653	50	7.7%
Virginia Tech	221	31	14.0%
Washington University in St. Luis	660	57	8.6%
Total & Average Response Rate	**3,467**	**255**	**8.4%**

Of the 255 survey respondents from our funded researcher group, the awards were then analyzed by their disciplinary category (Table 3); 49 percent of all participants had awards in the biomedical sciences, while another 20 percent of awards were classified as multidisciplinary. Our sample varied from our population ($x^2(5) = 7.23$, $p = 0.004$), with slight overrepresentation from environmental science and psychology and under representation from biomedical sciences. Only respondents who indicated they shared data from their grant were asked to provide expense information; 178 (69.8%) respondents said they shared data.

Table 3: Number and percent of Researcher Survey respondents, by award discipline area.

Award Discipline	Number of Researcher Responses, By Discipline	Percent of the Total Number of Responses, By Discipline	Percent of Responses From the Original Researcher Population, By Discipline
Multidiscipline	52	20.4%	18.8%
Biomedical Sciences	125	49.0%	59.4%
Environmental Science	19	7.5%	4.3%
Materials Science	12	4.7%	4.4%
Physics	16	6.3%	5.1%
Psychology	31	12.2%	8.0%
Response Total	**255**	**100%**	**100%**

Expense Data Cleaning

Of the 255 responses, only 91 researchers (36%) provided expense information. This high nonresponse rate for the expense questions reflects the difficulty of assessing this information. Several researchers reached out to us expressing difficulty remembering or being able to pull out these costs for a specific grant, especially for those who have had many simultaneous grants from multiple agencies. The subset of data pertaining to reported expenses, such as infrastructure costs, salaries, etc., may have required clarification from the respondent. Responses that were unclear, such as those without annual or hourly demarcations, were clarified via email or, when applicable, during interviews.

Percentage time and number of staff hired for data management and sharing were combined into full-time equivalent (FTE) values by summing up the percent effort of each reported staff member. Salary and percent effort were also calculated to provide a total annual

cost associated with time supporting data management and sharing activities. These costs were calculated as a percentage of the total grant amount and in terms of annual grant costs. Because infrastructure costs were reported in bucketed categories, the staffing costs were added to the lower and upper bounds of the categories to create a range of combined costs. Infrastructure and total costs were also expressed as percentages of grant award amount for the upper and lower bounds separately. The upper bound was used for analysis.

Researcher Survey Limitations

The following are possible methodological limitations for the Researcher Survey:

- There was a high rate of nonresponse on the salary and expenses questions from researchers. We believe this is indicative of how difficult it is to distinguish and pinpoint the costs associated with data management and sharing within a grant, as some respondents volunteered that they did not know how to answer those questions.
- Grant disciplines were not included in the NSF award database API pull. After narrowing down other criteria (such as time period) the remaining awards were categorized manually, based on the grant names. As NSF funds many interdisciplinary projects, categorizing some of the grants was challenging; hence, some awards are categorized as multidisciplinary.
- Institutional emails were considered valid, but likely included researchers outside of our inclusion criteria, such as retired faculty. We only determined who was retired or no longer at their institution if they emailed us back directly informing us of their institutional status change, or if they asked us to remove them from the study.
- Questions 12–16 asked researchers how they, or their research team, engaged with data management and sharing activities. Most responses indicated doing at least one activity. However,

some responses indicated "No" to question 17, "Did you hire or rely on staff (including graduate and undergraduate students) for making the research outputs of this grant broadly available?" and/or did not provide a response for question 18, "Approximately how many staff did you hire or rely on to make the research outputs of this grant broadly available?" We can assume at least one person would be required to fulfill reported data management and sharing activities, even if the number of activities is minimal. Due to this, the salary questions (19 and 20) were likely not reported fully, causing our data to be underreported in these areas.

- During follow-up interviews with a sample size of researchers, we recognized a pattern for the salary and dedicated-time questions (17–21). As these questions asked about "staff" and "positions," it is likely many researchers did not include themselves (salary and time) in these responses. Using a broader term such as "entire research team" may be more inclusive, or simply asking respondents to include themselves in their responses. Therefore, data in these areas—percent of time toward the activities and salaries—is underreported.
- For the salary-focused questions, some respondents included benefits/fringe, while others did not.
- Questions regarding time dedicated to data sharing activities and salaries omitted asking researchers about how many years staff/the research team were funded; asking this would have helped distinguish between annual costs and per-grant costs. Our results likely underestimate labor costs that occur regularly over multiple years of a grant, as we took the salaries reported as the total cost over the entire grant, rather than ones that could (and likely did) repeat over multiple years of the grant award period.
- Although we asked researchers to provide their institutional academic department, we did not ask them to provide disciplines for their awarded projects. If we had done so, we would have been able to confirm our manual NSF award categorizations.
- Asking researchers to provide information on the kind of data collected or produced during their awarded projects, including

data sizes and types, would have been beneficial in expense-data analysis. Analyzing similar data types and sizes, and their expenses, might yield more comparable results than comparing within disciplines only.

- Although we did not see a difference in terms of nonresponse by grant year, researchers who reported on grants from early in our 10-year timeframe (such as 2013–2016) may have underreported which activities they utilized for data sharing. Expenses for salaries and infrastructure may also be more likely to be underreported for these older grants. Surveying researchers on recently completed projects would likely yield more accurate results.

Interview Methodology

The last question on both the Institutional Infrastructure Survey for administrators and the Researcher Survey asked participants if they would be interested in a follow-up interview to provide further detail for their survey responses. In the administrator group, 49 of the total 69 survey respondents (80.3%) selected yes to a follow-up interview. In the researcher group, 32 of the total 255 survey respondents (12.6%) selected yes to a follow-up interview. Due to time restrictions, not all respondents could be interviewed; criteria were developed to select who, from these two groups, would be contacted to participate in a follow-up interview.

Administrator criteria for interview selection included:

- Potential clarification of expense data, as reported in the original survey response
- Representation of at least two people from each of the four service areas (IT, LIB, RSCH, IC) and at least two people from each institution

Researcher criteria for interview selection included:

- Potential clarification of expense data, as reported in the original survey response
- Representation from each of the discipline areas, with at least two interviews per discipline, and at least two people from each institution

In total, 12 researchers (two from each institution) and 15 administrators (at least two from each institution) were interviewed. Interview question templates (see Research Instruments #3 and #4 below) were developed for each group. The templates were structured to allow modifications of the questions based on the information participants provided in their survey responses. Of the 15 administrator interviews, 2 were guided interviews that occurred after the close of the survey. One participant requested an interview, as they did not make the survey deadline, and the second participant was approached directly by a RADS PI, as her response represented the library and, as such, the

RADS team deemed representation from the libraries as critical to our study. In this format, participants were asked the survey questions in addition to several questions from the interview template instruments. Questions from the interview instrument typically pertained to the expense questions, as well as future investment questions (non-retrospective).

All interviews were conducted remotely via Zoom and were scheduled for 60 minutes, although actual interview times ranged from 25 to 65 minutes. All interviews occurred between January 20 and March 27, 2023. The RADS project manager was present for all interviews and supported the six RADS PIs in interviewing administrators and researchers from their respective institutions.

The survey respondents who agreed to be interviewed were encouraged to invite team members (either from their unit or lab) engaged in data sharing work to participate in the interview. Interviews consisted of one to six participants, and participation depended on the survey respondent's identification of those in their unit or lab engaged in data sharing activities. At their discretion, survey respondents invited others to participate in the interviews.

Transcript Cleaning

All interview transcript editing was performed by the RADS project manager and occurred between January 30 and May 5, 2023. Transcripts were edited to:

- Correct words and phrases
- Correct speaker attribution
- Remove duplicate words when they appeared consecutively
- Fully spell out acronyms, especially when they referred to institutional units or infrastructure; common acronyms such as NIH or NSF were not spelled out
- Reflect position titles instead of names, except when directly referring to part of the RADS research team

Interview Coding Methodology

A subset of the RADS research team was responsible for the coding of researcher and administration interview transcripts. The first step in developing the codebook involved the qualitative research subgroup defining a set of guiding questions. Once a set of guiding questions was agreed upon, each team member independently coded two researcher and two administrator transcripts with categories they deemed relevant to the guiding questions.

As a group, the qualitative coding subgroup reviewed each team member's codes, adjudicated on them, and finalized a set of codes, definitions, and examples for each code they determined was relevant to the guiding questions. From these codes, a codebook with hierarchical codes was developed for full coding of all interview transcripts. See Appendix C for the codebook.

Each member of the qualitative coding subgroup was then tasked with coding 13 or 14 interviews. Each transcript was coded by at least two members of the qualitative coding subgroup and, to limit bias, no member of the subgroup coded interviews they facilitated. Two of the subgroup members used NVivo software to code the interviews while two other members used Taguette software. Once coding was completed, subgroup members uploaded their coded files into Google Drive. For those who used Taguette, a student assistant at one of the RADS institutions transformed the Taguette files into NVivo coded files.

Strategies to Improve Responses

Several lessons were learned in our methodology to increase participation in both the surveys and interview. They are noted below to provide insight for others who may want to duplicate or modify our research processes:

- **For administrative units, consider using a two-phased open survey approach.** Tableau visualizations[13] highlighting data sharing activities reported by the administrators from the institutional units who completed the survey were produced as a result of the research. In order to increase responses, it is recommended that a first draft of the visualization be sent to nonrespondents. This tactic may increase survey participation because many administrators will want to see their unit represented on a campus-wide scan.
- **Narrow the scope of the surveys.** We recognize that our research sought to address multiple research questions, and that others may only wish to use a portion of these methods. For instance, if the goal is to gain insight into an institution's research data management and sharing infrastructure, we recommend omitting the expense questions, as this may have been a barrier for some respondents.
- **Employ a guided-interview methodology.** While the survey instruments provide a good deal of guidance in terms of describing data management and sharing activities, we recognize that the interviews were ideal situations to provide ample context about the RADS Initiative. The interviews were also ideal spaces for respondents to freely ask questions about the survey and interview questions.
- **Gain project buy-in from high-level administration.** Buy-in from the Office of the Vice President for Research (OVPR) or similar offices will demonstrate the importance of understanding

13 For examples, see the following visualizations for: Cornell University, Duke University, University of Michigan, University of Minnesota, Virginia Tech, and Washington University in St. Louis (gathered at https://public.tableau.com/app/profile/cynthia.vitale8121/vizzes).

the institution's data sharing practices to the entire institution. This may be especially relevant as investments to support these services and activities are increasing at many institutions.

Research Instruments

The following research instruments were used for this research:

- Research Instrument #1 - Institutional Infrastructure (Administrator) Survey
- Research Instrument #2 - Researcher Perspectives Survey
- Research Instrument #3 - Administrator Interview Template
- Research Instrument #4 - Researcher Interview Template

Data Availability Statement

De-identified response data and data dictionaries for both the Institutional Infrastructure and Researcher surveys are located in the Washington University in St. Louis WashU Research Data (WURD) repository, at https://doi.org/10.7936/6RXS-103654.

Scripts for the federal funder API pulls are located in GitHub: https://github.com/DataCurationNetwork/rads-api-pulls/

Appendix A—RADS Data Management and Sharing Activities for Public Access

Version 1 of the RADS Public Access Data Management and Sharing (DMS) Activities was used for both the Institutional Infrastructure and Researcher surveys. The following are the DMS activities listed by participant group, and grouped by data life-cycle phase.

Planning, Design, and Start Up of Projects Phase	
Access Data Management and Sharing Activities - Institutions	**Public Access Data Management and Sharing Activities - Researchcers**
Reviewing or preparing data management plans (DMPs) or data management and sharing (DMS) plans	Preparing data management plans (DMPs) or data management and sharing (DMS) plans
Reviewing data management and sharing costs and expenses to be included in grant budgets	Identifying data management and sharing costs to be included in grant budgets
Reviewing of institutional review board (IRB) protocols and informed consent for data sharing	Preparing institutional review board (IRB) protocols and informed consent for data sharing
Developing, building, or recommending storage solutions for active research data	Determining storage solutions for active research data
Supporting an appropriate repository (or repositories) for making research data broadly available	Selecting an appropriate repository (or repositories) for making research data broadly available
Assessing data security needs and recommending solutions	Evaluating data security needs
Supporting intellectual property and copyright considerations	Determining intellectual property and copyright considerations
Developing or reviewing materials transfer agreements and/or data use agreements (DUAs)	Developing materials transfer agreements and/or data use agreements (DUAs)
Referring to disciplinary or institutional standards and/or best practices for handling, collecting, and documenting data	Reviewing disciplinary or institutional standards and/or best practices for handling, collecting, and documenting data

Data Collection, Storage, and Management Phase	
Public Access Data Management and Sharing Activities - Institutions	**Public Access Data Management and Sharing Activities - Researchcers**
Developing or reviewing documentation of data (for example, data dictionary, protocols)	Developing documentation of data (for example, data dictionary, protocols)
Creating quality-control mechanisms or procedures	Creating quality-control mechanisms or procedures
Evaluating or recommending data-analysis tools and processes to support sharing and reproducibility	Evaluating data-analysis tools and processes to support sharing and reproducibility
Managing active data (for example, storage, security, backup, lab notebooks)	Managing active data (for example, storage, security, backup, lab notebooks)

Making Data Broadly Available Phase	
Public Access Data Management and Sharing Activities - Institutions	**Public Access Data Management and Sharing Activities - Researchcers**
Consulting on decisions about what data to share or host	Making decisions about what data to share or host
Providing or hosting repositories for making data available	-
Preparing or consulting on preparing data for sharing (for example, de-identification, check privacy/personally identifiable information (PII)/protected health information (PHI), selection, curation, data cleaning, validation, and quality control)	Preparing data for sharing (for example, de-identification, check privacy/personally identifiable information (PII)/protected health information (PHI), selection, curation, data cleaning, validation, and quality control)
Submitting data into a data sharing platform (for example, institutional repository, generalist repository, disciplinary repository)	Submitting data into a data sharing platform (for example, institutional repository, generalist repository, disciplinary repository)
Creating or reviewing documentation for research data (for example, structured metadata, README files)	Creating documentation for research data (for example, structured metadata, README files)

Consulting, selecting, or applying licenses to data	Selecting or applying licenses to data
Recommending or migrating data file formats to be open or more accessible	Migrating data file formats to be more open or accessible
Creating or recommending persistent identifiers (PIDs; for example, digital object identifiers (DOIs))	Creating persistent identifiers (PIDs; for example, DOIs)
Checking for compliance with existing data use agreements (DUAs)	Checking for compliance with any existing data use agreements (DUAs)

Data Retention, Including Preservation, Archive, and Long-Term Access Phase	
Public Access Data Management and Sharing Activities - Institutions	**Public Access Data Management and Sharing Activities - Researchcers**
Consulting on or migrating files to new formats or systems as needed	Migrating files to new formats or systems as needed
Monitoring integrity of preserved data	Monitoring integrity of preserved data
Making decisions about de-accessioning and removal of research data	Making decisions about de-accessioning and removal of research data
Ensuring data security when appropriate (for example, PHI/Health Insurance Portability and Accountability Act (HIPAA), export controls, Federal Information Security Management Act (FISMA), student data, and intellectual property)	Ensuring data security when appropriate (for example, PHI/Health Insurance Portability and Accountability Act (HIPAA), export controls, Federal Information Security Management Act (FISMA), student data, and intellectual property)

Project Closeout and Compliance Phase	
Public Access Data Management and Sharing Activities - Institutions	**Public Access Data Management and Sharing Activities - Researchcers**
Ensuring funding agency requirements for data sharing have been met	Ensuring funding agency requirements for data sharing have been met
Providing compliance support around research project reports	Providing compliance support around research project reports

Note: The RADS Public Access DMS Activities were revised in December 2023. The updated version is online at https://doi.org/10.29242/radsdmsactivities2023.

Appendix B—Administrative Unit Categorization

The following is a list of all responding units/departments from the Institutional Infrastructure Survey, and their categorization, used in the project Tableau visualizations.

Institution	Responsive Department/Office	Service-Area Categorization
Cornell University	Center for Advanced Computing	IT
	Center for Technology Licensing	RSCH
	College of Engineering/Bowers College of Computing and Information Science (CIS)/Tech/IT Service Group (ITSG)	RSCH
	College Research Office	RSCH
	Cornell Center for Materials Research (CCMR)	IC
	Cornell Center for Social Sciences	IC
	Cornell Institute of Biotechnology	IC
	Cornell IT (CIT)	IT
	Cornell University Library	LIB
	Information Security Office	IT
	Research Development/Dean's Office	RSCH
	Sponsored Programs in the College of Life Sciences	RSCH
Duke University	Campus IRB	RSCH
	Duke Office of Research Initiatives	RSCH
	Medical Center Library	LIB
	Office of Campus Research Development (OCRD)	RSCH
	Office of Information Technology - Central IT Financial	IT

Institution	Responsive Department/Office	Service-Area Categorization
	Office of Information Technology - Central IT Operations	IT
	Office of Science Integrity	RSCH
	School of Nursing	IC
	University Libraries	LIB
University of Michigan	Innovation Partnerships	RSCH
	Information and Technology Services (ITS) - Advanced Research Computing (ARC)	IT
	Medical School Office of Research	RSCH
	Michigan Institute for Data Science (MIDAS)	IC
	Office of General Counsel	RSCH
	Office of Regulatory Affairs, Medical School	RSCH
	Office of Research - Innovation Partnerships	RSCH
	Office of Research UM - Flint	RSCH
	Office of Research and Sponsored Projects	RSCH
	Office of the Vice President for Research (OVPR) - Research Data Stewardship Initiative	RSCH
	Office of the Vice President for Research (OVPR) - Research Integrity	RSCH
	University of Michigan Biological Station (UMBS)	IC
	University of Michigan Library	LIB

Institution	Responsive Department/Office	Service-Area Categorization
University of Minnesota	Center for Transportation Studies	IC
	Chemical Engineering and Materials Science	IT
	Clinical and Translational Science Institute (CTSI)	IC
	College of Liberal Arts	IC
	Export Controls Office	RSCH
	Genomics Center (UMGC)	IC
	Health Sciences Technology	IT
	Masonic Cancer Center	IC
	Neuroscience/University Imaging Centers	IC
	Office of General Counsel	RSCH
	Office of Information Technology (OIT)	IT
	Office of Information Technology - University Information Security (OIT-UIS)	IT
	Office of the Vice President for Research - Risk Intelligence & Compliance Team (OVPR/RIACT)	RSCH
	Office of the Vice President for Research (OVPR) - Office of Biotechnology Activities Oversight	RSCH
	Office of the Vice President for Research (OVPR) - Technology Commercialization	RSCH
	Research Computing	IT
	University Archives-University of Minnesota Libraries	LIB
	University of Minnesota Libraries	LIB
Virginia Tech	Advanced Research Computing	IT

Institution	Responsive Department/Office	Service-Area Categorization
	Data Services - University Libraries	LIB
	Fralin Biomedical Research Institute at VTC	IC
	Information Technology Security Office and Lab	IT
	Office of Sponsored Programs	RSCH
	Research and Innovation	RSCH
	Virginia Tech Transportation Institute (VTTI)	IC
Washington University in St. Louis	Bernard Becker Medical Library	LIB
	Institute for Informatics	IC
	Office of the Chief Information Officer (OCIO) - Research Infrastructure Services	IT
	Office of the Vice Chancellor for Research	RSCH
	Office of the Vice Chancellor for Research - Joint Contracts and Research Development (JCRD)	RSCH
	Sponsored Projects Accounting & Office of Sponsered Research Services	RSCH
	University Libraries	LIB

Appendix C—Qualitative Data Codebook

The following codebooks were developed after our interviews with institutional administrators and researchers. These codebooks represent the systematic categorization and interpretation of key themes and insights extracted from the interview data.

Administrators — Realities of Academic Data Sharing (RADS) — Qualitative Analysis Code Book

Code_Parent	Code_Child	Description
Roles	Oversight/Supervision	The interviewee indicates who in the institution, lab, or elsewhere had responsibility for which part of the data sharing.
Roles	Other	
Responsibilities	Planning	Planning, Design, and Start Up of Projects
Responsibilities	Collection and Management	Data Collection, Storage, and Management
Responsibilities	Sharing	Making Data Broadly Available
Responsibilities	Retention	Data Retention, Including Preservation, Archive, and Long-Term Access
Responsibilities	Closeout	Project Closeout and Compliance
Tools	Repository	The interviewee indicates which repository or other tools they used to meet public access requirements
Tools	Data Management Plan	The interviewee indicates a data management plan was used to meet public access requirements
Public Access Preparations	NIH	The interviewee indicates how they have prepared or will prepare for the NIH DMS
Public Access Preparations	OSTP	The interviewee indicates how they have prepared or will prepare for the forthcoming OSTP-related public access policies

Administrators — Realities of Academic Data Sharing (RADS) — Qualitative Analysis Code Book

Code_Parent	Code_Child	Description
Barriers		The interviewee indicates a specific barrier, challenge, or burden brought on by the public access to data process.
Impact on Data Practices	Time	The interviewee indicates the amount of time (or change in time) data sharing requires
Impact on Data Practices	Documentation	The interviewee indicates the amount of documentation data sharing requires
Impact on Data Practices	Intellectual Property/ Ownership	The interviewee indicates intellectual property/ownership questions with data sharing
Impact on Data Practices	Training/Education	The interviewee indicates a change in training/education to meet data sharing requirements
Impact on Data Practices	Storage/Security	The interviewee indicates a change in storage/security needs to meet data sharing requirements
Costs	Budgeting	The interviewee indicates how budgeting was undertaken
Costs	Resources	The interviewee indicates which resources they leveraged to faciltate data sharing
Needs		The interviewee indicates a specific need they have to meet data sharing requirements that is not currently available

Researchers — Realities of Academic Data Sharing (RADS) — Qualitative Analysis Code Book		
Code: Parent	**Code: Child**	**Description**
Services		The interviewee indicates which campus-based services they used to meet public access requirements
Tools	Repository	The interviewee indicates which repository or other tools they used to meet public access requirements
Tools	Data Management Plan	The interviewee indicates a data management plan was used to meet public access requirements
Metadata Standards		The interviewee indicates which data or metadata standards they used to make data publicly accessible
Data Types		The interviewee indicates which data types they produced in their research
Costs	Budgeting	The interviewee indicates how budgeting was undertaken
Costs	Resources	The interviewee indicates which resources they leveraged to faciltate data sharing
Roles	Oversight/Supervision	The interviewee indicates who in the institution, lab, or elsewhere had responsibility for which part of the data sharing/research.
Roles	Lab Manager	The interviewee indicates who in the institution, lab, or elsewhere had responsibility for which part of the data sharing/research.
Roles	Research Team Member	The interviewee indicates who in the institution, lab, or elsewhere had responsibility for which part of the data sharing/research.
Responsibilities	Planning	Planning, Design, and Start Up of Projects
Responsibilities	Collection and Management	Data Collection, Storage, and Management
Responsibilities	Sharing	Making Data Broadly Available

Researchers — Realities of Academic Data Sharing (RADS) — Qualitative Analysis Code Book		
Code: Parent	**Code: Child**	**Description**
Responsibilities	Retention	Data Retention, Including Preservation, Archive, and Long-Term Access
Responsibilities	Closeout	Project Closeout and Compliance
Barriers		The interviewee indicates a specific barrier, challenge, or burden brought on by the public access to data process.
Impact on Data Practices	Time	The interviewee indicates the amount of time data sharing requires
Impact on Data Practices	Documentation	The interviewee indicates the amount of documentation data sharing requires
Impact on Data Practices	Intellectual Property/ Ownership	The interviewee indicates intellectual property/ownership questions with data sharing
Impact on Data Practices	Training/Education	The interviewee indicates a change in training/education to meet data sharing requirements
Impact on Data Practices	Storage/Security	The interviewee indicates a change in storage/security needs to meet data sharing requirements
Impact on Data Practices	Access/Reuse	The interviewee indicates a change in data access and reuse needs to meet data sharing requirements
Impact on Data Practices	Preservation	The interviewee indicates a change in data preservation needs to meet data sharing requirements
Needs		The interviewee indicates a specific need they have to meet data sharing requirements that is not currently available
Disciplinary Practice		The interviewee indicates disciplinary practices for data sharing