The Cost and Price of Public Access to Research Data: A Synthesis

“Scientific data underlying peer-reviewed scholarly publications resulting from federally funded research should be made freely available and publicly accessible by default at the time of publication, unless subject to limitations (...). Federal agencies should develop approaches and timelines for sharing other federally funded scientific data that are not associated with peer-reviewed scholarly publications.” (Office of Science and Technology Policy, 2022)

“...federal agencies should allow researchers to include reasonable publication costs and costs associated with submission, curation, management of data, and special handling instructions as allowable expenses in all research budgets.” (Office of Science and Technology Policy, 2022)

[**Introduction 2**](#_rzee9iioxo6)

[“Ensuring Free, Immediate, and Equitable Public Access” 3](#_skpafqu2ymit)

[**Key Concepts and Models 4**](#_uqi8iwmrus33)

[Cost and Price of Publication 4](#_itnx7vmcym7c)

[Allowable and Reasonable Costs 5](#_fbi76pxwysdf)

[**Public access to research data 7**](#_kyxsjl5zbk1)

[Publishing models for research data 7](#_xpzfd6vfsvu)

[The cost of providing public access to research data 9](#_j4iy9yply0kj)

[The price of providing public access to research data 15](#_ia40ya3gpz6j)

[Generalizability of previous research and implications for public access policy compliance 19](#_8im9qfpx2ix5)

[**Conclusions 21**](#_rv9k1hi1pgzw)

[**Acknowledgements 22**](#_wm14u2em242g)

[**References 22**](#_9jk64ak5c0l5)

# 

# Introduction

Beginning by or before December 31, 2025, all recipients of US federal research funding will be required to make their federally-funded, peer-reviewed scholarly publications freely available including scientific data, via public access venues (“agency-designated repositories”[[1]](#footnote-0)) with no delays or embargos after publication.

Issued by the US Office of Science and Technology Policy (OSTP), and signed by Director Alondra Nelson, the *Memorandum on Ensuring Free, Immediate and Equitable Access to Federally Funded Research* (Office of Science and Technology Policy 2022; also known as the “Nelson memo”) extends the reach of federal public access policy that previously was established in the 2013 *Memorandum on Increasing Access to the Results of Federally Funded Research* (Office of Science and Technology Policy 2013; also known as the “Holdren memo”). These extensions include: 1) revoking all embargos or delays in favor of immediate free access; 2) requiring all federal agencies to participate, not just those with R&D budgets of $100M or more; 3) requiring agencies to develop or extend policies on scientific data sharing to include all data from funded research, not just that which is directly related to publications; 4) ensuring the collection of specific types of metadata for all scholarly publications and data at the time of their deposit; and 5) advancing concerns related to equity of both participation in research and using the results (e.g., using assistive technologies). These significant changes are intended to increase and advance equity, American scientific leadership, and public trust in government funded science.

This white paper explores the spectrum and blend of cost and pricing models and options currently used to make research data free and publicly accessible. We begin by exploring the context for the Nelson Memo, including what “reasonable” and “allowable” costs currently look like within the US federal agency landscape.

This paper then identifies and synthesizes what is currently known about publishing research data in terms of **cost** (defined here as what value/resource is used up in order to produce, deliver, and maintain a research product) and **price** (defined here as the amount that is charged to or paid by stakeholders in the market in exchange for the service of producing, delivering, and maintaining a research product). Notably, both costs and prices exist for research data, even when they are not easily visible and regardless of who ultimately pays for them. As we show below, good information on both cost and price is hard to find and incentives for publishers and repositories of different types to tally, let alone provide this information, are largely missing.

This white paper has been undertaken by the NSF-funded “Investigating “reasonable costs” to achieve public access to federally funded research and scientific data” (2023-2025) project team based at Invest in Open Infrastructure,[[2]](#footnote-1) a not-for-profit entity that works to improve funding and resourcing for open technologies and systems supporting research and scholarship. With it, we intend to chronicle part of the problem space we are trying to address and understand in our own research. A companion white paper on the cost and price of publishing articles will be issued by our team this spring.

## “Ensuring Free, Immediate, and Equitable Public Access”

The Nelson memo, issued in August 2022, sought to ensure that all federally funded research (“publications and their supporting data”) be made available through “free, immediate, and equitable public access” by December 31, 2025. It also asked agencies to promote discoverability and accountability through the use of appropriate persistent identifiers.

Ideas about how to craft and implement policies following the Nelson Memo’s “free, immediate, and equitable” aim arose immediately after its publication, and these ideas demonstrate the divides and disagreement between different stakeholders.[[3]](#footnote-2) The implications of the forthcoming policies have been hotly discussed and debated over the last year, including through a range of interagency meetings, many of which include non-government stakeholders from each of the core communities impacted by the Nelson Memo.

The key aim outlined in the Nelson Memo is to make the benefits and advances in scientific research that have been supported by US taxpayers as available as possible to all Americans. This is in close keeping with the 2016 Principles for Promoting Access (Interagency Working Group on Open Data Sharing Policy (National Science and Technology Council), 2016); and is congruent with international trends for both national and private funders that have been strengthening since the NIH issued its first public access policy in 2005 (National Institutes of Health, 2005). The work of expanding “public access” can be accomplished through a variety of business models and mechanisms. Importantly, the Memo does not specify a government-run system as the end goal, although PMC and other agency-specific repositories demonstrate that this could be a part of the ecosystem that ultimately supports implementation.

# Key Concepts and Models

## Cost and Price of Publication

As many have noted, even when digital research outputs are free to access for users, the publication and management of these resources are not “free.”[[4]](#footnote-3) This raises significant questions regarding who should pay and how much they should pay to enable the publication, dissemination, and curation of research outputs. While most of the debates, both nationally in the US and internationally, have centered on journal articles, the same general challenges apply for publishing data. No publisher or repository operates without expenses, and those expenses–including labor and infrastructure–have to be covered by some set(s) of stakeholders, whether those are institutions (research libraries, societies, government agencies), extramural funding (foundations, government sources), or individuals (authors, depositors, readers).

While the revenue models for journals and for data repositories differ, expanded public access policies do prompt questions regarding the “gap” between the true costs of publishing and the prices that might be charged to stakeholders, particularly with the potential to shift the burden of cost to researchers. This potential gap between cost and price is still far from clear for most players. The **cost** (what value/resource is used up in order to produce, deliver, and maintain a publication) to a publisher or a repository to publish a scholarly article or dataset has been hard to calculate at best, and fully occluded at worst, by the range of functions that could be considered part of the publishing process. The processes of data preparation (checking, organization, formatting), documentation, transfer, deposit and storage, rights clearance, etc., can be handled in a wide variety of ways and at a range of levels of diligence, leading to significant differences in the costs of deployment. Sunk costs and in-kind contributions can increase the complexity of tallying publishing costs. Further confounding the equation, these “costs” may also support and include expenses well beyond the publication. As one example, societies often engage in fee-based article publishing in order to fund other mission-driven operations, including graduate student stipends, research awards, and other forms of discipline-based support to scholars.

The **price**, what is charged to or paid by stakeholders in the market in exchange for the service of providing public access to a research output–producing, delivering, and maintaining an article or dataset online–can be difficult to pin down for both articles and for datasets. The most direct expression of “price” for the service of providing public access to a dataset is a deposit fee, but as we shall see, a variety of revenue models are used by data repositories and may include institutional subsidies (usually from the host, and sometimes from extramural funders) or membership fees, making it difficult to assess whether a given price is reasonable in relation to the true cost of the service.

Our research herein seeks to determine what is known currently about 1) how much publishing (providing public access to) research data costs, and 2) what prices are charged for publishing research data. We undertook this research as a complement to other recent studies, including the OSTP Report to the U.S. Congress on Financing Mechanisms for Open Access Publishing of Federally Funded Research (Office of Science and Technology Policy, 2023).

While this work focuses primarily on the US context, we acknowledge and point to the reciprocal influences, such as plan S (cOAlition S, 2020), the UNESCO Recommendation on Open Science (UNESCO 2021), and similar efforts, that flow between different nations and regions. Likewise, while we are most concerned with understanding the implications of the 2022 Nelson memo for publishing models and mechanisms for federally funded research data in the US, we recognize that this is just one part of a sizable industry.

Finally, we have attempted to be clear and consistent throughout our analysis that, unless otherwise noted, when we use the term “cost,” it always references the resources used up in the process of publishing, and when we use the term “price,” it always refers to a monetary exchange between someone (author, reader, funder, institution) and the publisher. Notably, this does not fully map back to the Nelson Memo and its use of the terms “reasonable” and “allowable” costs.

## Allowable and Reasonable Costs

Underlying the implementation of the Nelson memo and the resulting agency policies are two complementary views related to cost. "Allowable" costs refer to the charges and activities that may be included in grant budgets, and the notion of "reasonable" costs aims to constrain those costs to some sensible amount. As noted above, we have tried to carefully distinguish between “cost” and “price” in our analysis, but for this section, we use the language of the Nelson memo and existing and emerging policies and refer to “costs” in this context as the amounts researchers might include in their grant budgets that they can then use to pay the prices associated with publishing their data and/or articles.

What, then, are the “allowable costs” that researchers might include in grant proposal budgets for their research data outputs?

While agency policies are still in early-if-active development, we look to the policy and planning documents of NIH and NSF for early insight into what two major federal funders might consider in scope in terms of allowable costs. The NSF (National Science Foundation, 2023b) takes a fairly expansive approach to allowable costs for data, including:

…cleanup, documentation, storage and indexing of data and databases; development, documentation and debugging of software; and storage, preservation, documentation, indexing, etc., of physical specimens, collections or fabricated items. Line G.2. of the proposal budget also may be used to request funding for data deposit and data curation costs.

The NIH released an updated data management and sharing policy that took effect in January 2020 (National Institutes of Health, 2020a), extending its reach to all NIH agencies and seeking more details on scientists’ plans for sharing, with the aim of increasing data sharing (Kaiser, 2019). The policy contains exclusions, including institutional overhead, as well the “costs of doing research,” which include those “associated with collecting or otherwise gaining access to research data (e.g., data access fees).” The policy deems allowable those costs associated with curating and documenting data, unique “local data management considerations,” and preserving and sharing data via existing repositories (National Institutes of Health, 2020b). While there is some ambiguity as to which of these costs are more appropriately allocated to the costs of doing research or to data sharing, overall the policy seems to point towards data curation activities that directly support preparation of data for sharing.[[5]](#footnote-4) All existing U.S. policies require that funds allocated for data sharing be spent during the period of performance of the grant award.

The Association of Research Libraries Realities of Academic Data Sharing (RADS) team interviewed representatives of the Department of Energy, Department of Transportation, Institute of Museum and Library Services, and Department of Agriculture, asking them to define allowable and non-allowable expenses, and to distinguish between the activities that are associated with “good scientific practice” and data management and sharing. Almost everything was considered allowable (the exception being proposal and project development). Agency representatives were also not overly concerned with establishing boundaries between the practice of good science and data management and sharing are not clear, as long as expenses are not charged more than once (Taylor and Narlock, 2024).

In terms of reasonable costs, the Nelson memo also advises agencies to work, *in consultation with OMB* (the Office of Management and Budget), to allow researchers to include such costs in their budgets. Section 200.404 of the Code of Federal Regulations (2 CFR 200.404 -- Reasonable Costs, 2023) defines a cost as "reasonable if, in its nature and amount, it does not exceed that which would be incurred by a prudent person under the circumstances prevailing at the time the decision was made to incur the cost."

For a concrete example of guidance offered around “reasonable costs,” we might look to travel policies and caps set for travel funded by the US federal government. These include per diem spending for different locations, limits on tiers of pricing for flights and ground transportation options, firm guidance on entertainment and alcohol expenditures, and other such guidance on what a federal grant can be used to fund or reimburse. Were OMB to move to develop analogous restrictions or limitations on the type and amount of publication fees that can be covered by a federal grant, clearer evidence and documentation regarding the cost and price of publication, and their variance, would be necessary.

Having clarified our use of the terms “cost” and “price”, as well as “reasonable and allowable costs,” we focus the rest of this paper on our analysis of the available information on the cost and price of providing public access to research data.

# Public access to research data

In this section we describe some of the pathways research data take to final publication, and summarize some of the extensive body of research on the component activities that must be taken into account in a complete life cycle analysis of the costs of research data curation and sharing. In the process, we look at cost modeling experimentation in the fields of research data management and digital preservation to consider what might be relevant from their approaches.

## Publishing models for research data

The pathways to publication for research data are diverse and differ from those for article publications. They include publication of data to various types of repositories, publication of data as a supplement to a research paper, publication of a data paper, and making data available online via unique, dedicated infrastructure. Our working definitions of these broad categories or pathways are presented in Table 1.

TABLE 1. Descriptions and examples of pathways to publication for research data.

| **Publication pathway** | **Definition** | **Example(s)** |
| --- | --- | --- |
| Repository: Disciplinary | Repositories accepting data from specific disciplines. | ICPSR,[[6]](#footnote-5)  Worldwide Protein Databank[[7]](#footnote-6) |
| Repository: Generalist | Repositories accepting deposits from the research public and not serving one or more specific disciplines. | Figshare,[[8]](#footnote-7) Zenodo,[[9]](#footnote-8) Dryad[[10]](#footnote-9) |
| Repository: Institutional | Repositories hosted by research institutions, primarily (but not always exclusively) serving the deposit needs of their researchers. | Merritt (University of California),[[11]](#footnote-10)  Harvard Dataverse,[[12]](#footnote-11)  Chiba University Repository for Access To Outcomes from Research[[13]](#footnote-12) |
| Repository: Project-specific | Repositories dedicated to the output of specific projects or facilities. | NASA Distributed Active Archive Centers (DAACs),[[14]](#footnote-13)  CERN Data Centre[[15]](#footnote-14) |
| Research paper supplement | Publisher-hosted supplements to the related paper (hosting via an established repository may also be an option). | IOPscience,[[16]](#footnote-15)  CellPress,[[17]](#footnote-16)  American J. Psychiatry[[18]](#footnote-17) |
| Data paper | Generally brief publications, published in data journals, that describe a dataset. The dataset is usually (but not always) hosted separately from the paper. | Scientific Data (Nature),[[19]](#footnote-18)  Earth System Science Data,[[20]](#footnote-19)  Journal of Open Humanities Data[[21]](#footnote-20) |

The Nelson memo (Nelson, 2022) directs agencies to recommend or require that researchers use existing and appropriate online repositories, without requiring the use of *specific* repositories. Accordingly, we focus our discussion on the disciplinary, generalist, institutional and project-specific repository pathways. As a result, we exclude “Research paper supplement” from our analysis. A recent statement by Brooks Hanson, the American Geophysical Union’s executive vice president for science, also argues against publishing research data as research paper supplements (alongside the underlying research paper and within the journal the paper is published), noting a recent decision by a group of earth sciences journal editors to discontinue the practice (National Academies of Sciences, Engineering, and Medicine et al., 2023). Further emphasizing the role of established repositories as the preferred solution for data sharing, the National Science and Technology Council (NSTC) has enumerated the “desirable” characteristics of digital data repositories ([National Science and Technology Council](https://www.zotero.org/google-docs/?ORAFG3), 2022), and the Nelson memo prompts agencies to bring their repository selection criteria into alignment with those of the NSTC as much as possible. Data papers, brief publications that describe a dataset, are generally peer reviewed, appear in indexed data journals, and are citable in the same manner as conventional publications, typically (but not always) describe a dataset hosted in an external repository, apart from the data paper (Jiao et al., 2023; Walters, 2020). As such, the costs of publication can be considered along with those of conventional publications, while the externally hosted datasets they describe are nearly always accounted for in the repository pathways.

These four repository types differ in the deposit restrictions they may impose. Disciplinary repositories, as the name implies, accept data from researchers across a discipline. Generalist repositories, on the other hand, tend to be discipline agnostic. In both cases, (disciplinary and generalist repositories), if they operate on a membership model, affiliation with a member institution may be required, but institutional affiliation alone is not generally a criterion for deposit acceptance. In both cases, a fee for deposit may be charged. In contrast, institutional repositories usually require that depositors be affiliated with the institution hosting the repository. Finally, project-specific repositories are generally open only to researchers depositing materials associated with particular large scale research projects or instruments. Deposits to these last two types of repositories tend to be free of charge.

As we explore previous and current research on the costs and prices associated with providing public access to research data, we will focus our attention on repository-based publishing models and the activities that support deposit to repositories.

## The cost of providing public access to research data

Current and previous work on the costs of research data curation specifically, and digital curation and preservation more broadly,[[22]](#footnote-21) has generally approached the question from one of three points of view: 1) that of institutions and individual researchers planning for the costs of data curation, 2) that of repositories seeking to understand and manage their operational costs, and 3) that of funders looking to define the total cost of providing access to the research outputs of all funded projects. Below we summarize some of the relevant research on each of these approaches.

### Institution and researcher focused approaches

Institutions and communities of practice have sought to support their researchers with information and tools that enable researchers to plan for the costs of data management and sharing, and to understand the total cost to research institutions. At both levels, that of an individual researcher and their institution, these approaches seek to quantify the costs of activities associated with research data management and sharing across the entire research life cycle.

At the institutional level and explicitly in anticipation of new public access sharing policies for federal research, two important and current bodies of work have emerged that attempt to identify the range of participants and their roles in these processes, and to enumerate the costs to institutions of activities associated with providing public access to research data.

First, the Council on Governmental Relations (COGR), published an analysis of the cost to research institutions of complying with NIH’s data management sharing policy (COGR, 2023). Their cost analysis is based on survey responses from 34 institutions and appears to be oriented towards institutionally provided services, rather than the services of external providers such as generalist or domain repositories. Nevertheless, it provides useful estimates that might be considered upper bounds on the potential costs of compliance for institutions. Five activities of “potential burden” are identified, only one of which (“Data plan,” for drafting a data management plan to accompany a proposal) does not directly touch on the cost of data publication. Cost drivers factoring into their model were new staff, opportunity cost (reallocation of staff), IT, and training. Taking into consideration reported salaries and other cost rates, and distribution of effort across campus units and all activity areas except “Data plan,” COGR estimated an annual cost of just slightly over $1 million per year for institutions with more than $100 million in annual R&D expenditures. This approach makes sense for COGR’s purpose, that of demonstrating the total impact of public access requirements on research on institutions. It is likely too high an estimate for our analysis, as not every component of the defined areas of potential burden would be considered directly related to the cost of publication.

The Association of Research Libraries (ARL) and the Data Curation Network (DCN) also approach the issue of the impact of public access requirements on research institutions. With support from the Institute of Museum and Library Services (IMLS), they initiated a deep examination of data sharing via their project, the “Realities of Academic Data Sharing (RADS) Initiative” (RADS Initiative, n.d.); NSF grant awarded in August of 2021). The project team also notes close collaboration with COGR. The first phase of their work involved developing a life cycle view of data sharing activities, and determining who (researchers or institutional service providers) was participating in each of the activities at the project’s institutional partners (RADS Initiative, n.d.; Taylor et al., 2022). They have also published a preliminary analysis of DataCite metadata from each of the participating institutions, in order to determine where researchers are sharing data (Mohr, 2023). Early findings show that the distribution of activities varies across researchers and institutional service providers as well as partnering institutions (RADS Initiative, n.d.; Webb, 2023). The focus of this research has been primarily on understanding cost areas attributable to researcher and institutional effort of data sharing across the entire research life cycle. To date, the project has published no quantitative conclusions about costs.

Institutions that support researchers and research data management have developed numerous tools to support individual researchers and research teams in planning for the cost of data curation in research projects. Some address the full research life cycle (e.g. University of Bristol Research Data Service, 2021; Utrecht University, n.d.), while others focus more directly on costs associated with sharing (e.g. UK Data Service, 2022). Still others primarily support data management planning, but still outline the major activities that should be factored into any cost analysis; the DMPTool being one such example (UC Curation Center (California Digital Library), n.d.-b, n.d.-a). In Table 2 we summarize the typical activities identified in each of these sources that researchers are advised to consider in their planning process. We suggest that many of these activities would be required simply to complete the proposed research, and thus are not uniquely applicable to meeting public access requirements. Of those that are uniquely applicable to providing public access, we note that most (documentation, data formatting for deposit, deposit to repository, and rights clearance) most likely incur costs via human effort rather than infrastructure costs. Storage and backup, transfer and security is the exception, and depending on the circumstances, may in fact be more appropriately allocated to the cost of doing research. Either way, we submit the primary cost driver is likely the human effort required for these activities.

TABLE 2. Cost components from selected sources providing guidance to researchers for data management planning. Our assessments of which activities are potentially properly allocated to providing public access (that is, activities that would not otherwise be necessary in order to simply perform the proposed research) are indicated.

|  | **Source** | | | | |
| --- | --- | --- | --- | --- | --- |
| **Cost component** | **University of Bristol Data Service** | **Utrecht University** | **UK Data Service** | **DMP Tool** | **Unique relevance to data publication** |
| Participant consent | Y | N | Y | Y | N |
| Data description and cleaning | Y | Y | Y | Y | N |
| Documentation | Y | Y | Y | Y | Y |
| Digitization | Y | Y | Y | N | N |
| Data organization and formatting for research use | Y | Y | Y | Y | N |
| Data anonymization | Y | Y | Y | Y | N |
| Data formatting for deposit | Y | Y | Y | Y | Y |
| Transcription | Y | Y | Y | N | N |
| Storage and backup, transfer and security | Y | Y | Y | Y | Y |
| Deposit to repository | Y | Y | Y | Y | Y |
| Obtain existing data | N | Y | N | N | N |
| New data collection | N | Y | N | Y | N |
| Rights clearance | N | N | Y | Y | Y |

Note: Y= Yes; N=No

These approaches and guidance documents are intended to be generalizable and applicable across disciplines and publishing models, and do not result in quantitative estimates of the costs of providing public access to research data. The National Institute of Mental Health Data Archive cost calculator is unique in this regard in that the tool enumerates activities and includes cost calculations (with user-provided salary information) specific to the submission of data to a specific archive (NIMH Data Archive (NDA), n.d.). ​​The cost model is a full research data life cycle model, factoring in and scaling activities that would be considered a part of the cost of doing research, so that researchers can include sufficient resources in grant proposals. Activities associated primarily with the process of publication include administrative activities (reviewing the submission agreement, requesting NDA accounts, etc.), and data preparation, validation and submission. All costs (here, largely effort) are attributed to the research team and not the archive, so the model does not shed light on costs of the data as and after it is published by the NDA. Again, notably, cost categories and calculations emphasize the importance of human effort, rather than infrastructure.

One final example provides another approach to costing for both researchers and repositories. In order to provide budgeting guidance to researchers the Digital Endangered Languages and Musics Archiving Network (DELAMAN) (Digital Endangered Languages and Musics Archiving Network, 2014) explored two very different archival case studies and found strong overlap in the estimate cost ranges for curating their community’s data. Using the typical award amount that would produce the data collections examined in the two case studies, they concluded that 8% of total direct costs in research awards would support the costs of the services provided by the archives. The study advised researchers to include this amount in their grant proposal budgets.

### Repository-focused approaches

Motivated by the need for repositories of all kinds to remain operationally sustainable, a great deal of in-depth research exists on the repository activities and costs of curating and preserving digital content, much of it building upon the framework laid out in the Open Archival Information System (OAIS) reference model (Consultative Committee for Space Data Systems, 2002). Activity-based cost (“ABC”) modeling, applied in the repository context, first estimates the costs of resources and activities deployed in the delivery of a service, and then looks at likely expenditures in each resource and activity area in order to understand which resources and activities are the most important drivers of the cost of a service. Palaiologk et al.’s application of this approach at DANS (Data Archiving and Networked Services, a national data repository in The Netherlands) provides a useful and frequently cited illustration (Palaiologk et al., 2012). The DANS model was designed to quantify cost in terms of euros per dataset, in order to enable such calculations within a particular repository context. Other important variables in the work with DANS were research discipline and the complexity of datasets, but these still trail human effort in terms of importance in driving overall costs.

Many models have utilized this approach, or variations of it, including the projects Keeping Research Data Safe (KRDS) (Charles Beagrie Ltd., n.d.), which looked at the costs of data preservation at UK universities, and work by the Consortium of European Social Science Data Archives (CESSDA) (Beagrie & CESSDA, 2017), among others (see 4C Project, 2013 and Open Planets Foundation, n.d. for extensive lists of projects in this area). Of these studies, only the KRDS project drew general conclusions about how costs tend to be allocated, apportioning approximately 55% to outreach, acquisition and ingest, 31% to access, 9% to “other” and 5% to preservation and storage. While they did not clearly define these partitions, the KRDS project demonstrated that staff are the most significant cost overall, and noted that while the costs of preservation and storage are continuous, they do tend to decline over time (Beagrie, 2017), and this is consistent with the DANS findings.

Work published by the Royal Society (The Royal Society, 2012) posits a tiered model for data according to scale, value and importance, and then explores costs at repositories in each of the tiers. Repositories representative of each tier were asked for information on service provision, count and total volume of data, deposit and download activity, budget, and staffing. Like KRDS and others, the Royal Society found staffing to be the most significant cost in every case examined. At the time, the Worldwide Protein Data Bank reported that $6-7 million of their annual costs of $11-12 million were attributable to deposit and curation. This may be largely attributable to labor costs, but more detail is needed. The UK Data Archives (UKDA) reported that staff costs constituted a higher proportion of their total budget at £2.43 million of £3.43 million (about 71%). Information for the UKDA indicated that periodic upgrades in infrastructure can cause those numbers to fluctuate substantially from year to year, and this variability likely applies to other repositories as well. Dryad reported staffing costs of approximately $300,000/year of a total budget of $350,000, or 86%.[[23]](#footnote-22) The proportion of costs accounted for by staffing was similarly high for institutional repositories, at 96% for ePrints Soton and 71% for DSpace@MIT.

The 4C (Collaboration to Clarify the Costs of Curation) project, coordinated by Jisc and which ran from 2013-2015, was unique in that investigators worked with repositories directly to collect information about the costs of curation. These data were included in the Cost Comparison Tool (CCT), a component of the Curation Costs Exchange (CCEx) platform. Organizations were invited to contribute structured cost data, which was then normalized by activities aligned with the OAIS reference model. The tool supported anonymized peer and global comparisons, and was available only to qualified, registered users. The concerns potential data providers had around participation and data sharing highlight the difficulty in collecting this kind of real world data on the costs of curation (Thirifays et al., 2014).

Taken together, this avenue of research suggests that repositories can meaningfully assess and understand their own cost drivers, which are strongly impacted by the human effort invested in curation activities. In fact, for the repositories in the studies reviewed here, effort is the most important cost component. However, even if the repositories included in these studies can be considered representative of their types (university repositories, national data and disciplinary data repositories), the importance of labor as a component of cost may not apply to other types of repositories. For example, if the labor associated with data curation activities (such as preparation and documentation of data for deposit) may fall to depositors using general purpose repositories, and less support for these activities is provided by generalist repository personnel, labor may be a less important driver of cost. Decisions about how much effort to invest and what activities to invest it in likely vary by repository and repository type, discipline, and type of data, making it difficult to make meaningful and quantitative generalizations about the costs of research data curation across a spectrum of repositories and datasets.

### Agency-wide approaches

Finally, we are aware of two creative attempts to quantify the cost of research data curation for the output of an entire funding agency’s research activities.

In a 2013 study, Plale and colleagues developed a very rough estimate of the total financial impact of the Holdren memo's requirement for research data sharing, for research funded by the National Science Foundation. They estimated the total number of papers supported by NSF funding using data from multiple sources and made some assumptions[[24]](#footnote-23) regarding the distribution of datasets and their size across disciplines. They then arrived at a total count of datasets and volume of data per unit time. From those totals, and some assumptions about system architecture, they arrived at a per gigabyte cost of about $5.56, $0.90 of which is allocated to storage and operations (the remainder is allocated to curation costs) for providing public access to data for 15 years (Plale et al., 2013). While the information informing the exercise is dated and the model simplistic, it is noteworthy for its effort to quantify costs at the scale of a federal agency with a large funding portfolio.

A comparison of the cost of research data curation with the total research budgets of two UK agencies in the 2012 report “Science as an open enterprise” (The Royal Society, 2012) offers another way to contextualize the costs of research data curation. The British Geological Survey, with an annual research budget of £30 million, spent £350,000 in support of the National Geoscience Data Centre, or a total of 1.2% of their total research budget. In contrast, the National Centre for Atmospheric Science (UK) research budget, with an annual research budget of £9 million, spent £1 million (11% of the research budget) to sustain the British Atmospheric Data Centre in the same year.

A good next step towards developing a more robust understanding of costs at the level of US federal agencies would be to quantify the published datasets funded by each agency, and where they are deposited. From there, as our understanding of repository costs improves, we can start to shed more light on the costs and prices associated with public access to research data.

## The price of providing public access to research data

In the costs of research data sharing section, we organized our thinking around the pathways used by researchers to share their data, and focused on the use of repositories as the pathway most likely to be utilized to meet public access requirements. We considered costs from the perspectives of research institutions and researchers, repositories, and funding agencies.

In this section, we briefly consider the range of business models and revenue sources employed by research data repositories of three broad types: institutional repositories (repositories accepting deposits from an institutions’ researchers), specialized repositories (project- or program-specific as well as disciplinary repositories), and generalist repositories (repositories accepting a wide range of datasets, regardless of the institutional affiliation of the depositor). We then summarize available information on charges for data sharing for each repository type, and suggest preliminary implications for public access policy compliance.

### Repository business models and revenue

Each of these broad types of repositories rely on diverse business models and sources of revenue to sustain their operations, including (but not limited to) structural support (from host institutions, research organizations, research funders), membership fees (which may be required for data access or deposit privileges for affiliated researchers, or used to provide support for repository operations), deposit fees, end-user fees for services that go beyond basic access to data, and grants or contracts (Dillo et al., 2017; Ember & Hanisch, 2013; Eschenfelder et al., 2022; OECD, 2017). Because public access mandates require that research outputs be made available at no charge to end users (Office of Science and Technology Policy, 2022), end-user fees for basic access do not represent a viable means for significant cost recovery in compliance with the policies, and we keep our focus here on examining the practice of repository charges to depositors or their institutions.

### Repository types and pricing

#### Institutional repositories

Institutional repositories, whether limited to research data or open to other kinds of research outputs, by definition rely primarily on structural support from the institutions they serve (e.g. (Lynch, 2003). While usually free to use, institutional repositories may impose limits on the size of individual files or datasets, or on the total allocation available for individual researchers or research groups. These limitations may be put in place simply due to the practical limitations of file upload over http, a desire to manage overall repository growth, storage and preservation requirements, or both. Selected examples of repository charges and limitations on deposits for several institutional repositories are presented in Table 3. We focused our consideration on services that support the publication of research data, although they may also accept other material. We excluded from consideration institutional services that support data storage and management for active research projects, and not publication and long term access. The examples include locally- and externally-hosted repositories running on a variety of repository platforms. For a more comprehensive view of institutional repository practices, we queried repository metadata in re3data, a global registry of research data repositories, on 6 November 2023. The results confirm that very few institutional repositories charge deposit fees, with only 7 of more than 800 institutional repositories in the registry listed as charging a fee for deposit (re3data.org, n.d.) and Table 3, below.

Table 3. Deposit platforms, limits, and charges for selected institutional repositories.

| **Repository** | **Platform** | **Deposit limits** | **Repository charges** | **Source(s)** |
| --- | --- | --- | --- | --- |
| Data Repository for U of M (DRUM),[[25]](#footnote-24) University of Minnesota | DSpace (locally hosted) | 5GB per file 50 GB per dataset | None | (University of Minnesota Libraries, n.d.) |
| eScholarship@UMassChan,[[26]](#footnote-25) University of Massachusetts Chan Medical School | DSpace (commercially hosted) | 15GB per deposit | None | (Lamar Soutter Library, 2023) |
| KiltHub,[[27]](#footnote-26) Carnegie Mellon University | Figshare (commercially hosted) | 5GB per file, 20GB per account | None | (Carnegie Mellon University Libraries, n.d.) |
| Merritt,[[28]](#footnote-27) University of California system | Custom application (locally hosted) | Unknown | None\* | (California Digital Library, n.d.) |
| Purdue University Research Repository (PURR),[[29]](#footnote-28) Purdue University | Hubzero (locally hosted) | Individual: 1GB Sponsored project: 10GB | None for specified limits; additional storage may be purchased | (Purdue University Library, n.d.) |
| WashU Research Data,[[30]](#footnote-29) Washington University in St. Louis | TIND RDM (commercially hosted) | 999GB per submission | None | (Washington University in St. Louis Libraries, n.d.; WURD Data Sharing and Curation Policy, n.d.) |

\* Campuses are charged annually for preservation storage, nominally $150/TB/year.

#### Specialized repositories

We use “specialized repositories” as an umbrella term to encompass project- or program-specific as well as disciplinary repositories. Project- or program-specific repositories typically serve large scale research collaborations, specific instruments, research facilities or laboratories, or dedicated research programs. They may be purpose-built, and use is typically limited to affiliated researchers. Disciplinary repositories serve the research communities of entire academic disciplines, and tend to be agnostic regarding institutional affiliation. The literature on pricing for deposit to these types of repositories is scant, and we turn again to the re3data research data repository registry for information on conditions of deposit for these repository types. The re3data editorial team identifies repositories for inclusion in the registry (directly or from user suggestions), thoroughly reviews and documents available information for each repository, and completes an entry in the re3data registry.

It is possible to ascertain from the registry some patterns in conditions that must be met for a researcher to upload data to different types of repositories. We queried the re3data registry on 6 November, 2023, selecting the repository provider type of “data provider,” in order to exclude pure metadata catalogs from consideration. Version 4.0 of the re3data repository metadata schema (not yet fully implemented in the registry) will allow the specification of a range of applicable repository types, including disciplinary, multidisciplinary, governmental, project-related, and other (defined as “neither institutional nor disciplinary”, see Strecker et al., 2023). Most of these types have yet to be applied to the current repository descriptions, and the assigned types we encountered were institutional, disciplinary, and other, and we explored fee and other upload (deposit) restrictions across these types.

Deposit fees appear to be rare across all repository types (Table 4). A total of just 23 of 2900 data provider repository entries (repositories hosting data) indicate a fee for deposit. Examples include the Archaeology Data Service, Bitbucket, and protocols.io. We note that some services do offer free deposit for smaller datasets (size limits vary), charging for deposits exceeding a specified limit, or charging researchers with grant funding or other resources, but still providing a basic, free option for deposit. This creates some uncertainty as to how repositories with tiered service models are represented in the registry.

A requirement for membership or affiliation is somewhat more common among those types. When “membership” refers to payment of fees for subscribing institutions, that can be considered a fee, although it is unlikely to be passed on to an institutions’ researchers. Libraries sometimes pay these membership fees, or they may be paid by other units that provide research or technology support within an institution. A registration requirement is common and unsurprising, as a user account may be necessary to access a repository’s full functionality. It is not clear what is meant by “other” upload restriction. We note that it is possible that repository entries are incomplete, and an earlier analysis of re3data metadata raised this concern (Kindling et al., 2017) but also notes that entries are reviewed by two members of the re3data editorial team.

Table 4. Upload restrictions by repository type from the re3data repository registry (re3data.org, n.d.).

| **Repository type** | **Repository count** | **Upload restriction:**  **Fee** | **Upload restriction:**  **Institutional membership** | **Upload restriction:**  **Registration** | **Upload restriction:**  **Other** |
| --- | --- | --- | --- | --- | --- |
| Institutional | 829 | 7 | 422 | 194 | 107 |
| Disciplinary | 2143 | 14 | 220 | 713 | 555 |
| Other | 323 | 6 | 23 | 104 | 80 |
| Total\* | 2900 | 23 | 620 | 866 | 675 |

\* A repository may appear in more than one of the repository type and upload restriction categories. Totals represent the result count for a given set of conditions for all repository types, and not the sum of each repository type.

#### Generalist repositories

Generalist repositories “store and preserve a wide variety of data types and research outputs and usually accept data regardless of the type, format, content, disciplinary focus, or research institution affiliation” (Barbosa et al., 2022). Following a successful pilot that demonstrated the need for generalist repositories for research data that are not a fit for discipline-specific repositories (NIH Office of Data Science Strategy, 2020) the National Institutes of Health launched the Generalist Repository Ecosystem Initiative in order to incorporate generalist repositories into the NIH “data ecosystem,” with the intention of fostering “consistent capabilities, services, metrics, and social infrastructure” among selected generalist repositories (National Institutes of Health, n.d.). We look at the practices of this selected set of repositories to understand what prices researchers, funders and institutions might be charged when fulfilling public access requirements for research data (Table 5). Within this small sample, five of seven are free to use as long as deposit limits are not exceeded.[[31]](#footnote-30) Dryad and Vivli are the exceptions, but we note that institutional pricing is available for both repositories. In the case where a subscribing institution pays a fee, it is not known whether they absorb the cost or pass fees on to their users.

Table 5. Data from (figshare, n.d.; Stall et al., 2023)

| **Repository** | **Deposit limits** | **Repository charges** |
| --- | --- | --- |
| Harvard Dataverse | 2.5GB per file  1TB per researcher | None |
| Dryad | 300GB per dataset via http  1TB mediated  No limit per researcher | Independent researchers: $150 per dataset  Institutional pricing available |
| Figshare | 20GB per free account  5TB per file | None up to 20GB  $875 per 250GB |
| Mendeley Data | 10GB per dataset | None |
| OSF | 5GB private projects  50GB public projects  5GB per file | None |
| Vivli | >1TB mediated | $4,000 for <500GB  $10,000 for >500GB.  Institutional pricing available |
| Zenodo | 50GB per dataset | None |

## Generalizability of previous research and implications for public access policy compliance

### Implications for researchers and their institutions

Federal agencies are directed by the Nelson memo to guide researchers towards the use of repositories meeting the criteria set forth in the National Science and Technology Council’s “Desirable Characteristics of Data Repositories for Federally Funded Research” (National Science and Technology Council, 2022). This provides researchers with flexibility and choice: many have access to an institutional repository, or they may take advantage of the availability of a wide range of generalist and specialized repositories, many without fees if data meet size or other restrictions. Institutions may elect to subscribe to commercial services or non-profit membership services, or to host and even develop their own services.

Numerous models and checklists exist that can inform researchers’ data management and sharing plans. These tools can support researchers’ needs to budget for the full range of data curation and sharing activities across the lifetime of a project. How researchers choose among their options, and whether price, repository services, visibility within scholarly communities, or other factors influence their choices are areas for further investigation.

More expansive public access requirements have the potential to drive an increase in demand for data sharing services, impacting the repository services institutions subscribe to or provide directly. Whether and how research institutions are planning to meet this demand also merits deeper investigation.

### Implications for repositories

As we discussed earlier, while the business models and revenue generating practices of data repositories are diverse, it appears they seldom include direct charges to researchers, and only sometimes rely on charges to institutions. This has important implications for the sustainability of data repositories. The threats to the viability of repositories are real and significant; (Strecker et al., 2023) reported that 191 (6.2%) of the 3069 repositories in the re3data registry at the time of their investigation had, and that the median age of repositories at the time of closure was only 12 years.

Repository-oriented approaches to cost modeling aid in understanding the chief costs and cost drivers in a particular repository context, and can in turn inform changes to a repository’s systems, services, and revenue management to promote sustainable operations. At both levels, the *labor* of data curation is usually the most significant cost driver. The level of curatorial activity required to support a researcher’s or repository’s objectives is not something that is readily generalizable *across* repositories, although individual repositories should be in a position to determine their average costs per dataset or other unit (e.g. size), as both the NIMH data archive and Digital Endangered Languages and Musics Archiving Network (and no doubt others) have done.

Where revenue is not directly linked to deposit activity, or the costs are obscured by business models that rely on institutional payments, repositories are at risk of being less well prepared to meet increased demand for their services. This suggests needs for more comprehensive and up to date data on the charging practices of repositories, and a better understanding of how they are planning for supporting the public access requirements their users will be trying to meet.

# Conclusions

The Nelson memo and emerging policies make clear that data repositories are the preferred solution for meeting public access requirements. Even so, two significant areas of ambiguity are apparent. First, many repositories leveraging sources of revenue that do not immediately scale up with increased deposits, sustainability is potentially an important concern for many repositories. Second, researchers and their institutions stand to benefit from having greater clarity as to the cost-generating activities that are *allowable* in grant budgets and what constitutes a *reasonable* amount to pay in meeting public access requirements.

*Repository sustainability*

The financial models supporting data repositories are diverse and often not directly related to usage - funders or institutions might provide structural support, research institutions might pay membership fees, and only occasionally are users charged for deposit. Unlike journal publishing, where the costs have historically been borne by readers or their proxies, the economics of data repositories appear to be only loosely connected to usage. Better and more complete information on the depositor charging practices, membership models, and other forms of financial support for data repositories would help shed light on these issues, as would a current understanding of repositories’ plans and concerns for supporting greater use as a result of the requirements.

*Allowable and reasonable costs*

Institutions are understandably concerned about the total costs of providing public access to research data, although some analyses appear to include costs that are properly associated with the conduct of research and not specifically with publication. The conversation around public access costs would benefit from a clearer definition and delineation of the activities that directly support providing public access (as opposed to the research process itself), how best to estimate them, and who pays them.

Labor is the most significant cost for repositories and data curation, particularly in support of ingest and access, although the actual cost of data curation in repositories varies by discipline, characteristics of data, and level of curatorial services. If "reasonable" cost is not readily generalizable, with allowable activities more clearly defined and greater transparency in repositories’ curation costs, researchers and funders could more easily evaluate whether deposit or membership fees, if charged, are reasonable.

*Impact on research and research budgets*

All parties need a better understanding of the cost of public access compliance. Unless funders’ research budgets grow, allowable costs will impact the amount of funding available for direct support for research. One could argue that public access stands to benefit researchers in several ways: their own data are better documented, safely stored, and accessible back to the research team, collaborators, and others, and they have easier access to others' data, but currently, little is known about the balance of the costs of public access compliance with the benefits it is expected to produce.

# Acknowledgements

We would also like to thank [Name and Name] for providing feedback to the complete draft of this report.

Funding acknowledgement:

# References

[*2 CFR 200.404—Reasonable costs.* (2023). https://www.ecfr.gov/current/title-2/part-200/section-200.404](https://www.zotero.org/google-docs/?ORAFG3)

[4C Project. (2013). *D3.1—Summary of Cost Models—4C Project*. https://www.4cproject.eu/summary-of-cost-models/](https://www.zotero.org/google-docs/?ORAFG3)

[Aiwuyor, J. (2022, August 25). ARL Celebrates Biden-Harris Administration’s Historic Policy to Make Federally Funded Research Immediately Available. *ARL News*. https://www.arl.org/news/arl-celebrates-biden-harris-administrations-historic-policy-to-make-federally-funded-research-immediately-available/](https://www.zotero.org/google-docs/?ORAFG3)

[Anderson, R. (2022, August 29). *A New OSTP Memo: Some Initial Observations and Questions*. The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2022/08/29/a-new-ostp-memo-some-initial-observations-and-questions/](https://www.zotero.org/google-docs/?ORAFG3)

[Association of American Universities. (2022, August 25). *AAU Statement on OSTP Decision to Make Federally Funded Research Publicly Available*. https://www.aau.edu/newsroom/press-releases/aau-statement-ostp-decision-make-federally-funded-research-publicly](https://www.zotero.org/google-docs/?ORAFG3)

[Barbosa, S., Gibson, J., Pfeiffer, N., & Van Gulick, A. (2022). *Introduction to Generalist Repositories for NIH Data Sharing*. https://datascience.nih.gov/sites/default/files/GREI-Webinar-1-Sept-2022-508.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[Beagrie, C. (2017). *CESSDA SaW Costs Factsheet*. Zenodo. https://doi.org/10.5281/zenodo.3662447](https://www.zotero.org/google-docs/?ORAFG3)

[Beagrie, C., & CESSDA. (2017). *CESSDA SaW Cost-Benefit Advocacy Toolkit*. Zenodo. https://doi.org/10.5281/zenodo.3662487](https://www.zotero.org/google-docs/?ORAFG3)

[California Digital Library. (n.d.). *Merritt Digital Preservation Repository Policies & User Guidelines*. California Digital Library. Retrieved November 5, 2023, from https://cdlib.org/services/uc3/merritt/merritt-policies-and-procedures/](https://www.zotero.org/google-docs/?ORAFG3)

[Carnegie Mellon University Libraries. (n.d.). *KiltHub Repository: FAQs*. Retrieved November 5, 2023, from https://guides.library.cmu.edu/kilthub/faqs](https://www.zotero.org/google-docs/?ORAFG3)

[Charles Beagrie Ltd. (n.d.). *Keeping Research Data Safe: Cost-benefit studies, tools, and methodologies focussing on long-lived data*. Retrieved October 11, 2023, from https://beagrie.com/krds](https://www.zotero.org/google-docs/?ORAFG3)

[Clarke & Esposito. (2022, August 29). *Zero Embargo*. https://www.ce-strategy.com/the-brief/zero-embargo/](https://www.zotero.org/google-docs/?ORAFG3)

[cOAlition S. (2020). *Plan S Principles | Plan S*. https://www.coalition-s.org/plan\_s\_principles/](https://www.zotero.org/google-docs/?ORAFG3)

[COGR. (2023). *Results from the COGR Survey on the Cost of Complying with the New NIH DMS Policy | Council on Governmental Relations*. Council on Governmental Relations (COGR). https://www.cogr.edu/sites/default/files/DMS\_Cost\_of\_Compl\_May11\_2023\_FINAL%20%281%29.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[Consultative Committee for Space Data Systems. (2002). *Reference model for an Open Archival Information System (OAIS).* CCSDS Secretariat. http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[DCC. (n.d.). *What is digital curation? | DCC*. Retrieved December 14, 2023, from https://www.dcc.ac.uk/about/digital-curation](https://www.zotero.org/google-docs/?ORAFG3)

[Digital Endangered Languages and Musics Archiving Network. (2014). Report of the DELAMAN Costing Case Study. *DELAMAN.Org*. http://hdl.handle.net/11122/6928](https://www.zotero.org/google-docs/?ORAFG3)

[Dillo, I., Treloar, A., Lusoli, W., Kupiainen, I., Asch, M., Diepenbroek, M., Hayashi, K., Noh, S.-Y., Oh, S. K., Skålin, R., Sithole, H., van Deventer, M., Golliez, A., Grolimund, P., Cudre- Mauroux, P., Dearry, A., Hodson, S., Neylon, C., Madalli, D. P., … Harrower, N. (2017). *Business models for sustainable research data repositories* (Vol. 47). OECD Publishing. https://doi.org/10.1787/302b12bb-en](https://www.zotero.org/google-docs/?ORAFG3)

[DPC. (n.d.). *What is digital preservation? - Digital Preservation Coalition*. Retrieved December 14, 2023, from https://www.dpconline.org/digipres/what-is-digipres](https://www.zotero.org/google-docs/?ORAFG3)

[Ember, C., & Hanisch, R. (2013). *Sustaining Domain Repositories for Digital Data: A White Paper*. https://doi.org/10.3886/SustainingDomainRepositoriesDigitalData](https://www.zotero.org/google-docs/?ORAFG3)

[Eschenfelder, K. R., Shankar, K., & Downey, G. (2022). The financial maintenance of social science data archives: Four case studies of long-term infrastructure work. *Journal of the Association for Information Science and Technology*, *73*(12), 1723–1740. https://doi.org/10.1002/asi.24691](https://www.zotero.org/google-docs/?ORAFG3)

figshare. (n.d.). figshare plus: Publish big datasets. Retrieved November 7, 2023, from https://knowledge.figshare.com/plus

[Interagency Working Group on Open Data Sharing Policy (National Science and Technology Council). (2016). *Principles for promoting access to federal government-supported scientific data and research findings through international scientific cooperation*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/iwgodsp\_principles\_0.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-023-02625-x](https://www.zotero.org/google-docs/?ORAFG3)

[Kaiser, J. (2019, November 11). *Why NIH is beefing up its data sharing rules after 16 years*. https://www.science.org/content/article/why-nih-beefing-its-data-sharing-rules-after-16-years](https://www.zotero.org/google-docs/?ORAFG3)

[Kindling, M., Pampel, H., Van De Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., & Scholze, F. (2017). The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine*, *23*(3/4). https://doi.org/10.1045/march2017-kindling](https://www.zotero.org/google-docs/?ORAFG3)

[Lamar Soutter Library. (2023). *eScholarship@UMassChan Data Deposit Policy*. https://repository.escholarship.umassmed.edu/pages/data-deposit-policy](https://www.zotero.org/google-docs/?ORAFG3)

[Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libraries and the Academy*, *3*(2), 327–336.](https://www.zotero.org/google-docs/?ORAFG3)

[Mohr, A. H. (2023). *RADS Metadata Analysis*. https://ajhmohr.github.io/rads\_metadata/](https://www.zotero.org/google-docs/?ORAFG3)

[National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, & National Academies of Sciences, Engineering, and Medicine. (2023). *Stakeholder Actions to Implement Open Scholarship: Proceedings of a Workshop-in Brief* (P. Whitacre, Ed.; p. 27133). National Academies Press. https://doi.org/10.17226/27133](https://www.zotero.org/google-docs/?ORAFG3)

[National Institutes of Health. (2005). *NOT-OD-05-022: Policy on Enhancing Public Access to Archived Publications Resulting from NIH-Funded Research*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-05-022.html](https://www.zotero.org/google-docs/?ORAFG3)

[National Institutes of Health. (2020a, October 29). *NOT-OD-21-013: Final NIH Policy for Data Management and Sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html](https://www.zotero.org/google-docs/?ORAFG3)

[National Institutes of Health. (2020b, October 29). *NOT-OD-21-015: Supplemental Information to the NIH Policy for Data Management and Sharing: Allowable Costs for Data Management and Sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-015.html](https://www.zotero.org/google-docs/?ORAFG3)

[National Institutes of Health. (2022). *7.9 Allowability of Costs/Activities*. https://grants.nih.gov/grants/policy/nihgps/html5/section\_7/7.9\_allowability\_of\_costs\_activities.htm](https://www.zotero.org/google-docs/?ORAFG3)

[National Science Foundation. (2023a). *NSF Public Access Plan 2.0: Ensuring Open, Immediate and Equitable Access to National Science Foundation Funded Research*. https://nsf-gov-resources.nsf.gov/2023-06/NSF23104.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[National Science Foundation. (2023b). *Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 23-1) | NSF - National Science Foundation*. https://new.nsf.gov/policies/pappg/23-1](https://www.zotero.org/google-docs/?ORAFG3)

[NIH Office of Data Science Strategy. (n.d.). *Generalist Repository Ecosystem Initiative*. Retrieved November 7, 2023, from https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative](https://www.zotero.org/google-docs/?ORAFG3)

[NIH Office of Data Science Strategy (ODSS). (2020). *Exploring a Generalist Repository for NIH-Funded Data*. https://datasciencedev.prod.acquia-sites.com/sites/default/files/GREI-Exploring-GREI-Document-508.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[NIMH Data Archive (NDA). (n.d.). *NDA Data Submission Cost Estimation Tool*. National Institute of Mental Health (NIMH). https://s3.amazonaws.com/stage.nimhda.org/Documents/NDA\_Data\_Submission\_Cost\_Estimation\_Tool.xlsx](https://www.zotero.org/google-docs/?ORAFG3)

[OECD. (2017). *Business models for sustainable research data repositories*. *47*. https://doi.org/10.1787/302b12bb-en](https://www.zotero.org/google-docs/?ORAFG3)

Office of Science and Technology Policy[. (2013, February 22). *Increasing Access to the Results of Federally Funded Scientific Research*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp\_public\_access\_memo\_2013.pdf](https://www.zotero.org/google-docs/?ORAFG3)

Office of Science and Technology Policy[. (2022, August 25). *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[Office of Science and Technology Policy. (2023). *Report to the U.S. Congress on Financing Mechanisms for Open Access Publishing of Federally Funded Research*. White House Office of Science and Technology Policy. https://www.whitehouse.gov/wp-content/uploads/2023/11/Open-Access-Publishing-of-Scientific-Research.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[Open Planets Foundation. (n.d.). *Costing Digital Preservation*. Retrieved October 11, 2023, from http://wiki.opf-labs.org/display/CDP/Home](https://www.zotero.org/google-docs/?ORAFG3)

[Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. *International Journal on Digital Libraries*, *12*(4), 195–214. https://doi.org/10.1007/s00799-012-0092-1](https://www.zotero.org/google-docs/?ORAFG3)

Patel, Vimal. (2022, August 26). *White House Pushes Journals to Drop Paywalls on Publicly Funded Research.* The New York Times. Retrieved January 22, 2024, from https://www.nytimes.com/2022/08/25/us/white-house-federally-funded-research-access.html

[Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature*, *473*(7347), Article 7347. https://doi.org/10.1038/473285a](https://www.zotero.org/google-docs/?ORAFG3)

[Plale, B., Kouper, I., Seiffert, K., & Konkiel, S. R. (2013). *Repository of NSF-funded Publications and Related Datasets: “Back of Envelope” Cost Estimate for 15 years*. https://scholarworks.iu.edu/dspace/handle/2022/16599](https://www.zotero.org/google-docs/?ORAFG3)

[Purdue University Library. (n.d.). *PURR - PURR Project Space Allocation and Pricing*. Retrieved November 5, 2023, from https://purr.purdue.edu/pricing](https://www.zotero.org/google-docs/?ORAFG3)

[RADS Initiative. (n.d.). Realities of Academic Data Sharing (RADS) Initiative. *Association of Research Libraries*. Retrieved August 29, 2023, from https://www.arl.org/realities-of-academic-data-sharing-rads-initiative/](https://www.zotero.org/google-docs/?ORAFG3)

[re3data.org. (n.d.). *Registry of Research Data Repositories*. DataCite. https://www.re3data.org/](https://www.zotero.org/google-docs/?ORAFG3)

[SAA. (n.d.). *SAA Dictionary: Digital curation*. Retrieved December 14, 2023, from https://dictionary.archivists.org/entry/digital-curation.html](https://www.zotero.org/google-docs/?ORAFG3)

Stall, S., Martone, M. E., Chandramouliswaran, I., Federer, L., Gautier, J., Gibson, J., Hahnel, M., Larkin, J., Pfeiffer, N., Sedora, B., Sim, I., Smith, T., Van Gulick, A. E., Walker, E., Wood, J., Zaringhalam, M., & Zigoni, A. (2023). Generalist Repository Comparison Chart. https://doi.org/10.5281/zenodo.7946938

[Strecker, D., Axtmann, A., Bertelmann, R., Cousijn, H., Elger, K., Ferguson, L. M., Fichtmüller, D., Jones, C., Lindenmann, I., Neidiger, C., Nguyen, T. B., Pal, J. K., Pampel, H., Petras, V., Schnepf, E., Semrau, A., Ulrich, R., Upmeier, A., Vierkant, P., … Wright, S. J. (2023). *Metadata Schema for the Description of Research Data Repositories: Version 4.0*. https://doi.org/10.48440/re3.014](https://www.zotero.org/google-docs/?ORAFG3)

[Strecker, D., Pampel, H., Schabinger, R., & Weisweiler, N. L. (2023). *Disappearing repositories—Taking an infrastructure perspective on the long-term availability of research data* (arXiv:2310.06712). arXiv. https://doi.org/10.48550/arXiv.2310.06712](https://www.zotero.org/google-docs/?ORAFG3)

[Taylor, S., Carlson, J., Herndon, J., Hofelich Mohr, A., Kozlowski, W., Moore, J., Petters, J., & Hudson Vitale, C. (2022). *Public Access Data Management and Sharing Activities for Academic Administration and Researchers*. Association of Research Libraries. https://doi.org/10.29242/report.rads2022](https://www.zotero.org/google-docs/?ORAFG3)

Taylor, S., Narlock, M. 2024. Conversations with US Federal Agency Representatives: Exploring Data Management and Sharing Expenses. Association of Research Libraries. https://www.arl.org/blog/conversations-with-us-federal-agency-representatives-exploring-data-management-and-sharing-expenses/

[The National Science and Technology Council. (2022). *Desirable Characteristics of Data Repositories for Federally Funded Research*. The National Science and Technology Council. https://doi.org/10.5479/10088/113528](https://www.zotero.org/google-docs/?ORAFG3)

[The Royal Society. (2012). *Science as an open enterprise*. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/](https://www.zotero.org/google-docs/?ORAFG3)

[Thirifays, A., Sisu, D., Davidson, J., Haage, K., Faria, L., Grootveld, M., Stokes, P., & Middleton, S. (2014). *D3.3 Curation Costs Exchange Framework*. 4C Project. https://doi.org/10.7207/4C-3.3](https://www.zotero.org/google-docs/?ORAFG3)

[UC Curation Center (California Digital Library). (n.d.-a). *DMPTool*. Retrieved October 11, 2023, from https://dmptool.org/](https://www.zotero.org/google-docs/?ORAFG3)

[UC Curation Center (California Digital Library). (n.d.-b). *Themes*. GitHub. Retrieved December 13, 2023, from https://github.com/CDLUC3/dmptool/wiki/Themes](https://www.zotero.org/google-docs/?ORAFG3)

[UK Data Service. (2022). *Data management costing tool and checklist*. UK Data Service. https://dam.ukdataservice.ac.uk/media/622368/costingtool.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[UNESCO Open Science Advisory Committee. (2021). *UNESCO Recommendation on Open Science | UNESCO*. https://www.unesco.org/en/open-science/about](https://www.zotero.org/google-docs/?ORAFG3)

[University of Bristol Research Data Service. (2021). *Anticipating the costs of research data management, version 1.6*. University of Bristol. https://www.bristol.ac.uk/media-library/sites/research/documents/lib-res-sup/rdm/guidance-documents/planning/Anticipating%20the%20costs%20of%20research%20data%20management%20v1.6.pdf](https://www.zotero.org/google-docs/?ORAFG3)

[University of Minnesota Libraries. (n.d.). *Data management good practices: Comparing data repository services*. Retrieved November 5, 2023, from https://libguides.umn.edu/c.php?g=1164012&p=8497440](https://www.zotero.org/google-docs/?ORAFG3)

[Utrecht University. (n.d.). *Costs of data management—Research Data Management Support—Utrecht University*. Retrieved October 4, 2023, from https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management](https://www.zotero.org/google-docs/?ORAFG3)

[Walters, W. H. (2020). *Data journals: Incentivizing data access and documentation within the scholarly communication system* (1). *33*(1), Article 1. https://doi.org/10.1629/uksg.510](https://www.zotero.org/google-docs/?ORAFG3)

[Washington University in St. Louis Libraries. (n.d.). *Managing your data: Data Sharing and Curation*. Retrieved November 5, 2023, from https://libguides.wustl.edu/c.php?g=47355&p=303438](https://www.zotero.org/google-docs/?ORAFG3)

[Webb, M. (2023). *Visualizations of public access data management and sharing activities, by each RADS institution*. Tableau Software. https://public.tableau.com/views/rads\_data\_support\_activities\_institutions/rads\_cu?%3Adisplay\_static\_image=y&%3AbootstrapWhenNotified=true&%3Aembed=true&%3Alanguage=en-US&:embed=y&:showVizHome=n&:apiID=host0#navType=0&navSrc=Parse](https://www.zotero.org/google-docs/?ORAFG3)

[*WURD Data Sharing and Curation Policy*. (n.d.). Retrieved November 5, 2023, from https://data.library.wustl.edu/pages/?page=policies&ln=en](https://www.zotero.org/google-docs/?ORAFG3)

Thank you for including me in the review, this was a pleasure to read and well defined in terms of scope and suggested steps forward.

One overarching thing of note: Towards the beginning it was a bit difficult to differentiate when the discussion was referring to articles and journal publications or data publications. While this definitely better as the paper went along, I wonder if there is a way to distinguish between publication of articles and open sharing of data.

One other thought is around the European and UK studies, these studies are so important in framing methods and developments over time, but somewhere I would suggest just an acknowledgement that there are many differences between the governing bodies, the perspectives, the history, etc. of the EU and the UK. (This is in and of itself would be interesting to look at one day….)

Kind regards, Reid

1. We note that “deposit” to a designated repository, in the context of these public access policies, is used to mean *either* deposit of a complete dataset and accompanying documentation, *or* deposit of metadata to a designated repository, as long as the metadata include a link to a publicly accessible copy of the dataset in its hosted location. [↑](#footnote-ref-0)
2. <https://investinopen.org/> [↑](#footnote-ref-1)
3. For just a taste of the debates, see statements and opinion pieces, e.g., “White House pushes journals to drop paywalls on publicly funded research” (Patel 2022), “ARL celebrates Biden-Harris administration’s historic policy to make federally funded research immediately available” (Aiwuyor 2022), “AAU statement on OSTP decision to make federally funded research publicly available” (Association of American Universities 2022), “Zero embargo” (Clarke & Esposito 2022), “A New OSTP Memo: Some Initial Observations and Questions” (Anderson 2022). [↑](#footnote-ref-2)
4. This is true for research articles and for datasets, and we use in this section the terms “publishing” and “publication” to refer to the services and activities required to provide public access to research outputs of both types. [↑](#footnote-ref-3)
5. As reported at the time, NIH’s 2020 policy changes were meant to address ways researchers were not sharing their data. This policy extended data management to all of NIH’s funding recipients (not just >$500K), and it shifted from requiring DMPs (Data Management Plans) to DMSPs (Data Management and Sharing Plans) with compliance/enforcement possible. See e.g. (Kaiser, 2019) [↑](#footnote-ref-4)
6. [https://www.icpsr.umich.edu/](https://www.icpsr.umich.edu/web/pages/) [↑](#footnote-ref-5)
7. <https://www.wwpdb.org/> [↑](#footnote-ref-6)
8. <https://figshare.com/> [↑](#footnote-ref-7)
9. <https://zenodo.org/> [↑](#footnote-ref-8)
10. <https://datadryad.org/> [↑](#footnote-ref-9)
11. <https://merritt.cdlib.org/> [↑](#footnote-ref-10)
12. <https://dataverse.harvard.edu/> [↑](#footnote-ref-11)
13. <https://opac.ll.chiba-u.jp/da/curator/?lang=1> [↑](#footnote-ref-12)
14. <https://www.earthdata.nasa.gov/eosdis/daacs> [↑](#footnote-ref-13)
15. <https://home.cern/science/computing/data-centre> [↑](#footnote-ref-14)
16. [https://publishingsupport.iopscience.iop.org/questions/supplementary-material-and-data-in-journal-articles](https://publishingsupport.iopscience.iop.org/questions/supplementary-material-and-data-in-journal-articles/) [↑](#footnote-ref-15)
17. <https://www.cell.com/supplemental-information> [↑](#footnote-ref-16)
18. <https://ajp.psychiatryonline.org/ajp_ifora> [↑](#footnote-ref-17)
19. <https://www.nature.com/sdata/oa> [↑](#footnote-ref-18)
20. <https://www.earth-system-science-data.net/> [↑](#footnote-ref-19)
21. <https://openhumanitiesdata.metajnl.com/> [↑](#footnote-ref-20)
22. The Society of American Archivists, defines digital curation as “the actions taken to select, manage, preserve, and add value to digital data throughout its lifecycle” (SAA, n.d.). The Digital Curation Centre, whose work is focused more specifically on research data, describes digital curation in much the same way, but with specific application to digital research data (DCC, n.d.). Digital preservation is defined by the Digital Preservation Coalition as “the series of managed activities necessary to ensure continued access to digital materials for as long as necessary ...(digital preservation) refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organisational change” (DPC, n.d.). Thus cost modeling research in the digital curation and preservation communities are potentially applicable to understanding some of the costs associated with providing access to research data. [↑](#footnote-ref-21)
23. Incidentally, around the same time as the Royal Society report, (Piwowar et al., 2011) reported approximate annual costs for Dryad of $400,000, when the archive contained approximately 10,000 datasets, suggesting a mean of about $40 per dataset per year. [↑](#footnote-ref-22)
24. Some of the authors' assumptions include: the number of NSF-funded papers generated per year (64,340, based on searches of selected databases for papers published 2011-2012), that papers are distributed evenly across NSF directorates and that each paper generates one dataset with an average size of 1, 10, or 100GB (depending on NSF directorate), and that the cost of curation is $150 per dataset (based upon the authors' experience). Assumptions are also made about storage costs and their change over time, periodic infrastructure upgrades, and other operational costs. [↑](#footnote-ref-23)
25. <https://conservancy.umn.edu/drum> [↑](#footnote-ref-24)
26. <https://repository.escholarship.umassmed.edu/> [↑](#footnote-ref-25)
27. <https://kilthub.cmu.edu/> [↑](#footnote-ref-26)
28. <https://merritt.cdlib.org/> [↑](#footnote-ref-27)
29. <https://purr.purdue.edu/> [↑](#footnote-ref-28)
30. <https://data.library.wustl.edu/?ln=en> [↑](#footnote-ref-29)
31. Dryad and Vivli are the exceptions, but we note that institutional pricing is available for both repositories. [↑](#footnote-ref-30)