Engineering Data Repositories and Open Science Compliance: A Guide for Engineering Faculty and Librarians

Adam Lindsley, Oregon State University

Adam Lindsley is the Engineering Librarian at Oregon State University. He teaches graduate research ethics, science/information literacy for undergraduates, and library research skills for both. Research interests include information literacy, data management, photogrammetry, pedagogy, and learning technology.

Dr. Shalini Ramachandran, Loyola Marymount University

Shalini Ramachandran is the Research and Instruction Librarian for STEM at Loyola Marymount University in Los Angeles. Her research and teaching interests include algorithmic bias, ethical AI, virtual reality for lab instruction, and open science.

Clara Llebot, Oregon State University
Sheree Fu, California State University, Los Angeles

Sheree Fu is the Engineering, Computer Science, and Technology Librarian at California State University, Los Angeles.

Engineering Data Repositories and Open Science Compliance: A Guide for Engineering Faculty and Librarians

Introduction

As engineering and data management specialist librarians, we advocate for the core values of open science, open access publishing, and open data that further accessibility, inclusivity, collaboration, transparency, innovation, and global impact in scientific research. We provide research support and open science guidance to engineering faculty and students and respond to questions about data sharing and data repositories. In the United States of America, many of our faculty receive federal funding for their research endeavors. To best support our engineering faculty, staff, and students, we stay updated on the changing landscape of United States federal guidelines on research data sharing. The following are some of the landmark publications and policy changes that impact our constituents.

In April 2022, the White House Office of Science and Technology Policy (OSTP) and the National Science and Technology Council (NSTC) published "Desirable Characteristics of Data Repositories for Federally Funded Research" [1], outlining a set of recommended features and qualities that are considered desirable for data repositories handling research data resulting from federally funded research. The document establishes a set of standards and guidelines to ensure that data resulting from federally funded projects is preserved in repositories that effectively manage and disseminate it.

On August 25, 2022, Dr. Alondra Nelson, then Acting Director of OSTP, issued a Memorandum [2] recommending that all federal agencies formulate new plans or update existing ones, outlining their approach to ensuring public access to peer-reviewed publications and the research data associated with federally funded studies. The Memorandum (hereafter referred to as the Nelson Memo) builds on an earlier OSTP Memorandum issued in 2013 by John Holdren [3], and emphasizes immediate access to federally funded research outcomes (without embargoes), and involves all federal agencies.

With the deadline for public access policies upcoming on December 31, 2025, it is important for the engineering community to evaluate the scope and policies of existing engineering research data repositories and how they fit NSTC's "Desirable Characteristics" criteria. It is particularly important to focus on discipline specific repositories. Discipline specific repositories are usually preferred to generalist repositories (e.g., NIH guidance on how to choose a repository [4]) because research may be more discoverable by others in the same field who customarily use these repositories for their work, and because the standards used by the repository, their metadata, and the support they offer can be tailored to the type of data they work with. To the

best of our knowledge, a resource for engineering that evaluates various engineering repositories for alignment with NSTC's "Desirable Characteristics" guidelines does not exist at this time.

In this study we investigate, through a literature search, current data deposit practices in the engineering practitioner and research communities. We also review existing guidance or mandates to help engineers find and choose a repository for their research data . We then look at a range of engineering-specific data repositories to evaluate their appropriateness for engineering research data deposition, using the NSTC's "Desirable Characteristics" criteria. Based on our investigation, we were able to shortlist 35 data repositories that we can recommend to engineering faculty and librarians.

Literature Review

Data Sharing Practices Among Engineering Researchers

The literature about data sharing practices specific to engineering researchers is scarce, but it is consistent with reported data sharing practices in other fields. The effect of policy on researchers' data sharing attitudes and practices is a prominent topic. Clear guidance from government funding bodies is identified by authors as an essential factor, although it does not always translate in willingness or expertise to share data effectively. Thus, articles that investigate data management practices, including data sharing, by researchers that are not affected by data sharing requirements from federal funders find that data sharing is residual and focused on sharing by request. For example, Kervin and Hedstrom [5], in a study published in 2012 that pre-dates the first OSTP Memorandum in the U.S., saw that those with limited funding from agencies with data sharing requirements lacked motivation for publicly disclosing their data. The study reveals that alternative funding sources not only fell short in promoting data sharing but, in certain instances, actively discouraged it. Wallis, Rolando, and Borgman [6], in 2013, investigated data sharing and reuse practices within the Center for Embedded Network Sensing (CENS), a National Science Foundation (NSF) Science and Technology Center, over a ten-year period. At that time only a few domain areas mandated data deposition by funders or journals, limited repositories were available for CENS research data, and sharing primarily occurred through interpersonal exchanges. The findings reveal that CENS researchers expressed willingness to share their data, but the majority of participants were only willing to share under certain conditions, which included getting credit, being allowed to retain initial publication rights, sharing only if the effort to do so was minimal, or if the requestor is known to the group, or if mandated by funder or journal policies. When asked how they shared data, researchers majoritarily described sharing only by request, or posting data to a website. The use of repositories was mentioned by about one quarter of the participants, and few participants could name a repository that could be appropriate for their data.

Another example is the article by Aleixandre-Benavent et al in 2020 [7], which outlines the results of a Spanish survey conducted under the Datasea project in early 2015. In 2015, the only data management requirements existing for Spanish research projects was an optional data management plan for researchers who wanted to opt for Horizon 2020 funding from the European Commission. The aim of the questionnaire was to assess the practices of Spanish health science researchers regarding the management and sharing of raw research data. Of the respondents, 28% did research in the area of physics and technology. The electronic questionnaire, completed by 1063 researchers, covered data creation and reuse, and data preservation. They found that legal concerns were the primary barrier to sharing (47.9%), followed by fear of their data being misused or misinterpreted (42.7%), and concerns about losing authorship (28.7%). Researchers believed that the organization's repository was the best place to preserve and share data, but they were also unaware of existing repositories and their requirements. Very similar concerns were described by Chowdhury, Boustany, and Walton [8]. They describe the results of a survey involving university researchers in the UK, France, and Turkey. The three countries were, at the time of the study, at very different moments of data sharing policy development and implementation. They found that researchers harbored concerns about data sharing, mostly related to ethics (67.5%), but also misuse and misinterpretation of data, and fear of losing the scientific edge. They also had a lack of understanding regarding the necessary steps for making data publicly accessible. The study underscored the need for substantial training and advocacy efforts to actualize the vision of widespread data sharing.

There have been explorations on research data management best practices in the U.S. context as well. Wiley [9] examined data management perspectives of aerospace, industrial and mechanical science engineering faculty affiliated with University of Illinois Urbana-Champaign (UIUC). The author conducted fourteen semi-structured interviews and analyzed them using a qualitative inductive coding method. This study built upon previous work and sought to explore how the elements of data management planning align with researchers' workflow, challenges, and awareness of research data services. Overall, the goal of this study was to gain a better understanding of these researcher's data management practices and enhance the research data services provided to faculty and research groups. Additionally, Cooper et al. [10] reported on the Ithaka S+R's Research Support Services Program, examining the evolving research needs of civil and environmental engineering scholars in the United States, aiming to enhance support services for them. Conducted in collaboration with 11 academic libraries and sponsored by the American Society of Civil Engineers (ASCE), the project drew insights from prominent scholars in the field. They suggested that despite facing challenges common to STEM disciplines, such as funding competitiveness and data management complexities, civil and environmental engineering offers unique opportunities for academic support providers due to its collaborative and interdisciplinary nature. However, outdated research infrastructures hinder innovation, necessitating interventions that capitalize on the field's strengths in fostering personal and collaborative relationships within academia and with industry. The report outlined the distinctive

research practices of civil and environmental engineering scholars and offered implications for various stakeholders, including academic libraries, universities, publishers, and research technology developers.

Works that investigate the perceptions of data sharing in contexts affected by data sharing policy find that the presence of clear data management guidance and policy improves the motivation to share data, although problems of resources and expertise still remain, and they advocate for the need of support to overcome these difficulties. For example, a recent paper by Parker from 2023 [11] addresses the implementation of the Research Data Management (RDM) Policy by the Tri-Agency Council of Canada, with a focus on Natural Sciences and Engineering Research Council of Canada (NSERC). It highlights the mandate for Canadian post-secondary institutions, including U15 research institutions, an association of 15 Canadian public research universities,to create Institutional RDM Strategies. A survey at a U15 research institution reveals engineering faculty's limited preparedness and willingness to widely share data, despite their awareness of open access (OA) practices. The study underscores the importance of RDM support and proposes a role for subject librarians in assisting faculty and educating students on RDM best practices.

The effect of scientific discipline on data sharing perceptions and attitudes has been described by Tenopir and collaborators. In their 2015 article [12], they describe how relevant differences regarding data sharing behaviors, perceived risks and barriers are not correlated with age or geographic provenance, but are related to discipline. For example, researchers that work with human subjects are less willing to share data than researchers of other disciplines. The most relevant work regarding the attitudes of engineering researchers is the article by Chowdhury et al. [13]. They investigate open research data practices in materials science and engineering through in-depth interviews with 13 researchers. The findings highlight the diverse nature of research data in this field, often customized for specific research focuses and necessitating detailed descriptions for potential reuse. The results reveal a lack of familiarity with modern data search methods, bilateral data sharing influenced by supervisors, and project funding policies. Identified obstacles to sharing include legal restrictions and the time needed for precise data description. The study suggests actions like researcher training and rewards to support open data initiatives.

Guidelines for Sharing Engineering Data

OSTP has set December 31, 2025 as the deadline for all agency public access policies for publications and data to be in effect. Several U.S. federal agencies have already published draft or final policies about the sharing of research data, even though most of these are not in effect, yet. Below we summarize data sharing guidance from plans published by U.S. federal agencies that are most often funders of engineering research projects. Given the topic of this article, we

focus mostly on their expectations regarding the repositories where data must be shared and preserved.

The National Institutes of Health (NIH) had been developing guidelines for a data management and sharing (DMS) policy for several years prior to the Nelson Memo. Five months after the release of the Nelson Memo, in January 2023, NIH implemented the final form of their DMS policy [14], aimed at promoting the sharing of scientific data to advance human health. Under this policy, NIH investigators are required to prospectively plan, submit a DMS plan, and comply with the approved plan for managing and sharing scientific data. This NIH policy includes guidance on how to choose a repository [4], and encourages researchers to select repositories that exemplify a set of characteristics, defined in the same document, and compatible with the NSTC "Desirable characteristics" document. NIH recommends depositing research data in discipline specific repositories, but it recognizes that often these are not available, and maintains a list of NIH supported generalist data repositories [15], and established the Generalist Repository Ecosystem Initiative (GREI), a program to establish cohesive capabilities, metrics and infrastructure across generalist repositories [16].

Other federal agencies have published plans that are still being reviewed, and that are not effective yet. For example, the Department of Energy (DOE), in its Public Access Plan [17] released in June 2023 prompts researchers to write a Data Management and Sharing Plan where they will describe, among other things, how data sharing will be maximized, and data repository selection. The DOE does not endorse any particular repository and recommends using repositories that are appropriate for the data type and discipline, that reflect relevant standards and community best practices for data and metadata, and that align with the Desirable Characteristics document. The National Science Foundation (NSF) published in February 2023 an updated version of their NSF Public Access plan [18], which is being reviewed after a request for information period. In the plan, NSF recognizes the important role of discipline specific repositories, and commits to continue providing resources to repositories to ensure appropriate data preservation and access over time. The plan endorses the criteria listed in the "Desirable Characteristics" document, although it recognizes that additional federal guidance in this area is necessary.

Similarly, the National Aeronautics and Space Administration (NASA) released a draft of their public access plan in February 2023 and asked for feedback from the NASA community. The document requires researchers to write data management plans regarding their data, metadata, software, software documentation, and associated data. NASA describes acceptable ways of archiving, preserving and sharing data, which include NASA archives such as NASA Technical Reports Server (NRTS) and clarifies that the use of existing databases or public repositories for archiving and preservation will be strongly encouraged. The draft expects specific data management guidance to be given by program managers, including expectations for planned

repositories. The only two characteristics that repositories must include, according to the plan, are the ability to provide persistent identifiers, and the ability to provide appropriate-term access.

Other agencies have not published new plans, and are revising their current ones. The Department of Transportation (DOT) is in the process of incorporating feedback that they got to their plan for Increasing Public Access to the Results of USDOT-Funded Transportation Research, which was published in 2015 after the Holdren Memo. The plan requires researchers to present Data Management Plans with information on how machine-readable data will be deposited in public repositories, where appropriate and available, but there isn't specific guidance on which repositories are acceptable.

Regarding journal policies, Wiley [19] explores the effect of data-sharing regulations within engineering journals. The author recognizes attributes linked with policy effectiveness, and gauges the influence on the prevalence of data sharing. The examination encompasses 28 journal publications spanning 2016-2017, revealing that 76% of engineering journals exhibit relatively weak data-sharing guidelines, 6% possess strong policies, and 14% do not address data sharing. Variables such as open access (OA) status and impact factor (IF) do not demonstrate a correlation with the strength of policies. The study underscores the necessity for standardization in data policies and delves into the motivations and obstacles in sharing research data within the field of engineering.

Overall, the literature captures the makeshift nature of data sharing practices over the years and reveals the need for the clear directives that were ultimately issued in the Nelson Memo of 2022. We were unable to find literature that touched on good practices for data sharing specifically for engineering research, and confirmed that the guidance on choosing repositories by U.S. funders relies on the general directives outlined in the NSTC "Desirable Characteristics" document. We expect that new and updated data sharing plans derived from the OSTP Nelson Memo will also have similar advice regarding repositories for engineering research. Exploring the suitability of engineering repositories to support the OSTP and NSTC guidelines will be beneficial for researchers, lowering barriers for data deposition and clarifying data sharing compliance for federally funded research.

Methods

To create an initial list of discipline specific repositories to include in this study we used the Registry of Research Data Repositories (re3data.org), a global registry of research data repositories from all academic disciplines. Repositories in re3data are tagged in a variety of different ways. We selected repositories tagged with the subject "Engineering Sciences" and then filtered by the repository type "disciplinary," yielding 193 potential sites to investigate out of a total of 693.

We then made an initial assessment of the repositories to ensure that we only kept those that a researcher working in the U.S. on engineering projects would be able to use to deposit their research data. We eliminated repositories that did not accept uploads (even if they did host datasets of interest to engineers), repositories that had location restrictions outside of North America, repositories that charged access fees, or repositories that were no longer functional.

After this first assessment, the remaining repositories were then evaluated against eight out of the total fourteen criteria from the NSTC guidelines [1], of which six comprise the Digital Object Management section, one is from Organizational Infrastructure (Free and Easy Access), and one from Technology (Authentication) See Table 1 for a description of each criteria as shown in the NSTC guidelines.

Table 1. Descriptions of NSTC criteria used in the second evaluation of the repository list

Criterion	Description
Unique Persistent Identifiers	The repository assigns a dataset a citable, unique persistent identifier (PID or DPI), such as a digital object identifier (DOI), to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The unique PID points to a persistent location that remains accessible even if the dataset is de-accessioned or no longer available.
Metadata	The repository ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets, using schema that are appropriate to, and ideally widely used across, the communities that the repository serves.
Curation and Quality Assurance	The repository provides or facilitates expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.
Broad and Measured Reuse	The repository ensures datasets are accompanied by metadata that describe terms of reuse and provides the ability to measure attribution, citation, and reuse of data (e.g., through assignment of adequate and openly accessible metadata and unique PIDs).
Common Format	The repository allows datasets and metadata to be accessed, downloaded, or exported from the repository in widely used, preferably non-proprietary, formats consistent with standards used in the disciplines the repository serves.
Provenance	The repository has mechanisms in place to record the origin, chain of custody, version control, and any other modifications to submitted datasets and metadata.

Authentication	The repository supports authentication of data submitters. The repository has technical capabilities that facilitate associating submitter PIDs with those assigned to their deposited digital objects, such as datasets.
Free and Easy Access	The repository provides broad, equitable, and maximally open access to datasets and their metadata free of charge in a timely manner after submission, consistent with legal and policy requirements related to maintaining privacy and confidentiality, Tribal and national data sovereignty, and protection of sensitive data.

The objective of this paper is to provide a preliminary assessment of repositories with respect to quality rather than conducting a comprehensive examination. Certain characteristics from the NSTC guidelines have been omitted from our analysis such as Risk Management, Retention Policy, Long-term Organizational Sustainability, Long-term Technical Sustainability, and Security and Integrity, due to the difficulty of verifying them through a simple website search. Further, we did not include the Clear Use Guidance criterion as the Broad and Measured Reuse criterion also provides metadata for use guidance. Researchers utilizing this paper as a reference should exercise due diligence and independently verify the quality of repositories to ensure their suitability for their specific needs.

The repositories in this second assessment were divided into equal parts among three of the authors, and each repository was evaluated individually. Given a lack of uniformity in descriptive language among repositories as related to the NSTC guidelines, we did our best to interpret the intent of each repository's documentation and policies, rather than requiring a strict 1:1 connection between the criteria and the documentation. If upload, metadata, authentication, or format requirements were not obvious from the repositories' documentation, we attempted to make accounts and upload items to force the repository to demonstrate these processes. If it was unclear whether or not a repository met the criteria, we investigated further as a group to make a determination. The fourth author reviewed the evaluated list for consistency and cross checked documentation on re3data.org against descriptions on the individual repository websites.

Lastly, we decided to recommend repositories based on the fulfillment of at least five out of the eight criteria. By setting this threshold, we aim to ensure that the repositories recommended possess a sufficient level of desirable characteristics to meet the needs of researchers and data users. This approach offers advantages in comprehensive evaluation, flexibility, and balanced assessment. By considering multiple criteria such as accessibility, documentation, and data integrity, we ensure a thorough assessment that minimizes the risk of overlooking crucial factors. Establishing a threshold of five out of eight criteria allows for flexibility in selection while maintaining rigor and accommodating variations in repository offerings. Each criterion represents a distinct dimension of repository quality, ensuring a balanced assessment that meets diverse researcher needs. Ultimately, this methodology prioritizes essential aspects of repository

quality, empowers researchers with reliable recommendations, and contributes to the advancement of open science research endeavors.

Results

From our initial list of 193 "engineering sciences" + "disciplinary" repositories, 38 met the criteria for the first assessment. Among these 38 repositories, 35 met the NSTC criteria for the second evaluation (again, they must have met at least five of eight assessment areas). These remaining 35 are reported in Table 2.

Table 2. Repositories, engineering subjects, and criteria met

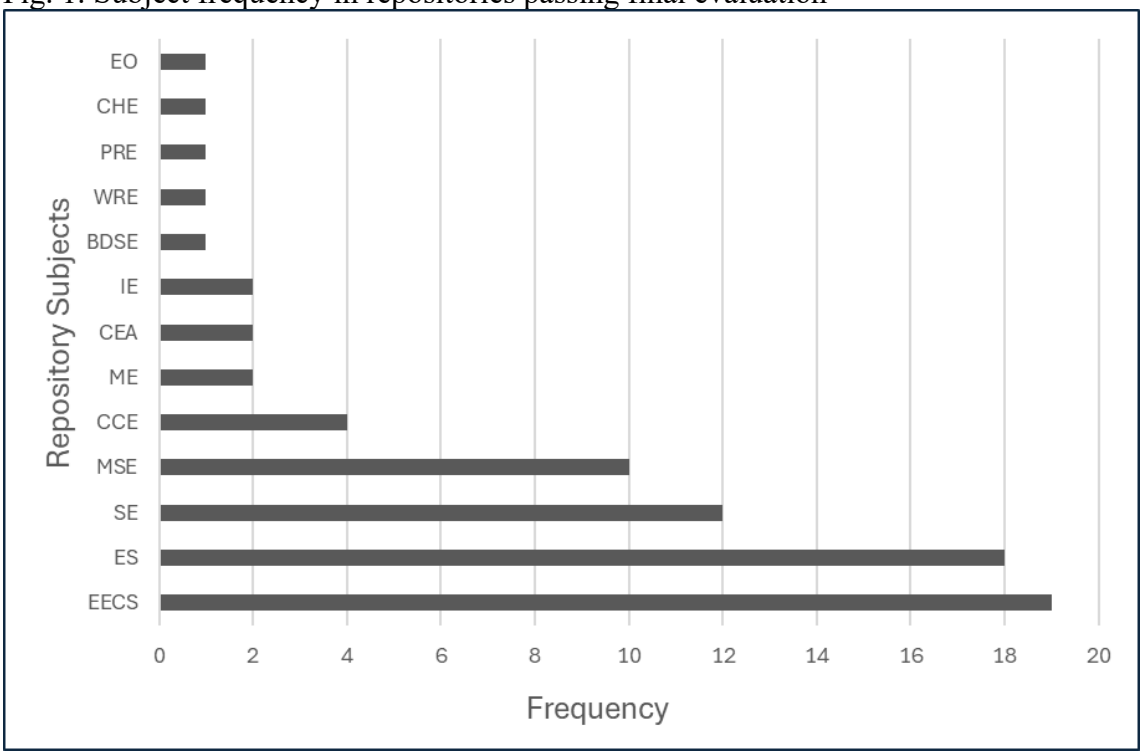
Repository Name	Criteria Met	Subjects
4TU.ResearchData | science.engineering.design	8	CCE
Buildings Data Platform	8	EECS, BDSE, CCE, CEA
DesignSafe-CI Data Depot Repository	8	CCE
Digital Rocks Portal	8	MSE, ES
Energy Data eXchange	8	EECS, SE
IMPACT	8	EECS, SE
MatDat.com	8	MSE, ME, IE
Materials Cloud Archive	8	MSE
Materials Data Facility	8	MSE
Materials Project	8	MSE, ES
NASA Socioeconomic Data and Applications Center	8	CEA, ES
NAWI Water DAMS	8	ES, WRE
OEDI	8	ES
Open Access Power-Grid Frequency Database	8	EECS, SE, PRE
OpenKIM	8	EECS, MSE, CHE, EO
Software Heritage Archive	8	EECS, SE
TInnGO Open Data Repository	8	EECS, SE, ES
UC Irvine Machine Learning Repository	8	EECS, SE, ES
Atmosphere to Electrons, Data Archive and Portal	7	EECS, ME, IE
Code Ocean	7	EECS
MHKDR	7	ES
National Science Digital Library	7	EECS, ES
NoMaD Repository & Archive	7	MSE, ES
Open Power System Data	7	EECS, SE, ES
OpenEI	7	ES
OpenML	7	EECS, SE, ES
QsarDB	7	MSE

The Perovskite Database Project	7	MSE, ES
VRP-REP	7	EECS, ES, SE
World Data Center for Renewable Resources and Environment	7	MSE, ES
brainlife	6	EECS, SE
nanoHUB	6	EECS, ES
Open Energy Platform	6	EECS, SE
Savannah	6	EECS, ES
SUITS Data Repository	6	EECS, SE, CCE

Reported subjects are limited to those that are engineering discipline-specific. Subject tag abbreviations: BDSE (Building Design Structural Engineering), CEA (Construction Engineering & Architecture), CCE (Civil & Construction Engineering), CHE (Chemical Engineering), EECS (Electrical Engineering & Computer Science), EO (Engineering Optics), ES (Engineering Sciences), IE (Industrial Engineering), ME (Mechanical Engineering), MSE (Materials Science & Engineering), PRE (Process Engineering), SE (Systems Engineering), WRE (Water Resources Engineering)

Considering the content of the repositories that passed our evaluation, nineteen have an Electrical Engineering and Computer Science (EECS) tag, ten had Materials Science and Engineering (MSE), and one was tagged both EECS and MSE. Generally, repositories with EECS tags also include Systems Engineering (SE) and/or Engineering Sciences (ES). The remaining six repositories' content consisted of a mix of mechanical engineering, construction engineering, chemical engineering/chemistry, water engineering, or used the general "engineering sciences" tag. Many also included non-engineering subjects, but these are not reported in Table 2. See Figure 1 for the frequency of subjects appearing in the final list.

Fig. 1. Subject frequency in repositories passing final evaluation



Of the final 35 repositories, 32 offer unique persistent identifiers, and three of them do not offer any unique identifier. DOIs are the most commonly used identifiers (26), but some of them also use handle identifiers. Two of the repositories use their own repository-specific unique identifiers.

Most repositories collect rich metadata about the datasets they host and have tools that help find data using discipline-specific metadata fields. In a few cases the metadata is more basic, and includes only the information needed to cite the dataset.

Curation and quality assurance are often mentioned in repository documentation, and many of the chosen repositories fulfill the curation criteria. It is important to note, though, that documentation rarely describes the actions, checks and criteria that are used during the curation process, so the curatorial effort likely varies greatly across the repositories in Table 2.

The majority of repositories have a statement about using common formats, but these are not necessarily open formats, as discipline-specific repositories have discipline-specific software and file formats. Most of the repositories (32) offer some manner of version tracking or provenance for uploaded datasets, though there is a wide variance in how this information is displayed to the end-user.

Most repositories supported a form of user authentication, connecting authors to datasets via ORCID or other persistent author identifier. However, many of these systems were voluntary on the part of the uploader and did not necessarily guarantee that the author or a designated representative was the person responsible for the dataset.

Finally, all but one of the repositories offer completely free download access to the datasets they host. The one exception is Code Ocean, which employs more of a platform model in which users run experiments with the datasets/code that have been uploaded, and are given a certain amount of processing time per month for free. Many repositories do require that download users make an account to access datasets.

The repository list we have generated, including both the initial and final evaluations, is available to download here: https://doi.org/10.7267/rf55zh75v

Discussion

Trends

During our analysis, we observed that repositories with a narrow focus on specific topics or data categories showed a dichotomy in their management approach. These repositories were either carefully documented and managed, often underpinned by grant funding or professional stewardship, or they were minimally documented and with content management conducted via direct communication with the site owner.

Most engineering repositories we examined that did not meet our evaluation criteria failed due to restrictive access, often bound by institutional affiliations, organizational memberships, or specific grant-awardee relationships. We observed that if a repository did have an open upload policy, it was then very likely to also meet the NSTC criteria; only three of the 38 repositories in the second evaluation did not pass. We did opt to incorporate a select few repositories with access limitations, as they were otherwise broadly inclusive; a prime example being OpenEI, a constituent of the repository ecosystem of the U.S. Department of Energy.

Considering multi-disciplinary repositories catering to engineering data, the options appear to be sparse. Our inquiry suggests that IEEE DataPort might be the most comprehensive, yet, overall, the engineering data repository landscape seems devoid of a unifying repository initiative such as the NIH's GREI. We did not include IEEE DataPort in our final list because it has restrictions on open access to deposited data. Up to two terabytes of cloud storage data can be uploaded on IEEE Dataport by researchers for free. However, the free uploads are not open data, and end users would need to be IEEE Dataport subscribers to access these datasets. There is an open access fee that researchers can pay to make their deposited datasets open, and those datasets are

then available to all users, including non-subscribers. Given this restrictive access model, we decided not to include IEEE Dataport.

For researchers contemplating contributing data to a repository, prior knowledge of the specific repository for their discipline is imperative. Our exploration highlighted re3data's broad interpretation of "engineering," a factor that may complicate discovery for users unfamiliar with the platform's tagging system or lacking knowledge of particular repository names.

We noted substantial representation of certain disciplines within these repositories, especially electrical engineering, systems, materials science, abandoned software tools, statistical analysis, remote sensing, energy, and machine learning. Conversely, the coverage of other engineering fields was distinctly thin. This disparity led us to wonder if the well-represented disciplines might be less susceptible to commercialization, or perhaps in these domains, the data itself does not constitute the commercially valuable component of the process. Alternatively, perhaps these fields have a deeper history of data sharing and are therefore inherently better represented in the repository ecosystem. We anticipate that the repository landscape, especially post-NSTC implementation, will undergo significant evolution. However, this evolution must be supported by funders; repositories will not be launched or improved without a substantial funding base.

Challenges

As librarians, we approach the integration of repository recommendations into our resources with a degree of caution and due diligence. Although we stand by our recommendations, there are several factors that should be considered before fully integrating this list into practice.

The impending implementation of the NSTC policy is likely to bring more clarity to the somewhat ambiguous documentation currently observed in repositories, particularly those engaged with state or federal agencies or receiving government funding. This should aid users in navigating and utilizing these repositories more effectively.

However, we observed a general lack of unity in the design and backend infrastructure of these repositories. While some adhere to a common infrastructure, many embed subject or discipline-specific terminology, complicating the user experience and navigation of their features. This diversity in design and language necessitates increased general repository knowledge for researchers/uploaders to overcome these potential barriers.

A notable concern is the ambiguity surrounding data donations. Several repositories solicit dataset donations without a clear definition of what constitutes a donation. This raises important questions: Does donating data equate to relinquishing all rights to it? Are donated datasets subjected to the same level of curation and review as standard submissions? Our impression is that these repositories have few datasets and are mostly distributing a very specific type of data

to their users; their primary goal is to be a data source, rather than a place where researchers can share their data.

The volatility of the repository landscape further complicates matters. During our project, approximately 10% of the repositories we monitored ceased operations, while others transitioned to commercial models. The dynamic nature of these platforms, including established commercial sites like IEEE DataPort, suggests that the decision to commit to a particular repository should not be taken lightly. In scenarios where a trustworthy, free repository is elusive, incorporating data sharing costs into grant applications, akin to publication fees, might be a more viable strategy.

Paying to publish data may be a solution to repository sustainability, especially after the initial period in which substantial funding will be going into repositories to respond to new policies. However, there is the danger that publishers could monetize access to these repositories, and it is important for the future of open science to avoid this outcome.

Assessing the longevity and stability of repositories poses a significant challenge for end-users. Indicators such as grant timelines or the commercial success of a repository can offer some insights, but they are not a guarantee. The question of sustainability and responsibility looms large, especially when funding for a repository ceases. What becomes of the datasets housed within these repositories? Are they fated to simply vanish, or should publishers take on the role of custodians, akin to HathiTrust's model for preserving out-of-print books? While governments may manage datasets in certain domains, there exists a gap for datasets that fall between these established categories, especially in the absence of dedicated funding or clear priorities.

Conclusion

In conclusion, the landscape of engineering data repositories reflects both progress and challenges. While there are commendable examples of effective repositories, the diversity and depth of engineering disciplines suggest a need for more robust offerings that align with NSTC’s Desirable Characteristics for data repositories. It is unfortunate that IEEE Dataport, the engineering repository with the broadest scope, operates under a restrictive paying model, limiting access and hindering the principles of open science. However, the guidance provided by NSTC, offers a framework for enhancing documentation and standardization across existing repositories. Looking forward, the OSTP recommendations arising from the Nelson Memo hold promise in incentivizing funding agencies to prioritize the development of new, discipline-specific repositories where needed. By leveraging existing and evolving resources and advocating for openness and collaboration, engineering librarians and data librarians can play a vital role in advancing open science compliance within the engineering community.

References

[1] White House Office of Science and Technology Policy (OSTP), "Desirable Characteristics of Data Repositories for Federally Funded Research," Executive Office of the President of the United States, May 2022. doi: 10.5479/10088/113528.

[2] A. Nelson, "Memorandum for the Heads of Executive Departments and Agencies: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research." Aug. 25, 2022. doi: 10.21949/1528361.

[3] J. P. Holdren, "Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research." Feb. 22, 2013. doi: 10.21949/1528360.

[4] National Institutes of Health, "NOT-OD-21-016: Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research." Accessed: Jan. 25, 2024. [Online]. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html

[5] K. Kervin and M. Hedstrom, "How research funding affects data sharing," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, in CSCW '12. New York, NY, USA: Association for Computing Machinery, Feb. 2012, pp. 131–134. doi: 10.1145/2141512.2141560.

[6] J. C. Wallis, E. Rolando, and C. L. Borgman, "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLOS ONE*, vol. 8, no. 7, p. e67332, Jul. 2013, doi: 10.1371/journal.pone.0067332.

[7] R. Aleixandre-Benavent, A. Vidal-Infer, A. Alonso-Arroyo, F. Peset, and A. F. Sapena, "Research data sharing in spain: Exploring determinants, practices, and perceptions," Data, vol. 5, no. 2, 2020, doi: 10.3390/data5020029.

[8] G. Chowdhury, J. Boustany, S. Kurbanolu, Y. Unal, and G. Walton, "Preparedness for research data sharing: A study of university researchers in three European countries," in *19th International Conference on Asia-Pacific Digital Libraries, ICADL 2017, November 13, 2017 - November 15, 2017*, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10647 LNCS. Bangkok, Thailand: Springer Verlag, 2017, pp. 104–116. doi: 10.1007/978-3-319-70232-2_9.

[9] C. Wiley, "Research Data Management: A Case Study Examining Aerospace, Industrial and Mechanical Science Engineering Faculty Research Practices," *Science & Technology Libraries*, vol. 42, no. 3, pp. 391–398, Jul. 2023, doi: 10.1080/0194262X.2022.2153780.

[10] D. Cooper et al., "Supporting the Changing Research Practices of Civil and Environmental Engineering Scholars," Ithaka S+R, Jan. 2019. doi: 10.18665/sr.310885.

[11] S. Parker, "Research Data Sharing in Engineering: A Report on Faculty Practices and Preferences Prior to the Tri-Agency Policy," in *2023 ASEE Annual Conference &*

Exposition Proceedings, Baltimore, Maryland: ASEE Conferences, Jun. 2023, p. 44112. doi: 10.18260/1-2--44112.

[12] C. Tenopir et al., "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide," *PLOS ONE*, vol. 10, no. 8, p. e0134826, Aug. 2015, doi: 10.1371/journal.pone.0134826.

[13] B. Suhr, J. Dungl, and A. Stocker, "Search, reuse and sharing of research data in materials science and engineering—A qualitative interview study," *PLOS ONE*, vol. 15, no. 9, p. e0239216, Sep. 2020, doi: 10.1371/journal.pone.0239216.

[14] National Institutes of Health, "NOT-OD-21-013: Final NIH Policy for Data Management and Sharing." Accessed: Feb. 24, 2023. [Online]. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

[15] National Institutes of Health, "Generalist Repositories." Accessed: Jan. 25, 2024. [Online]. Available: https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html

[16] National Institutes of Health, "Generalist Repository Ecosystem Initiative | Data Science at NIH." Accessed: Jan. 25, 2024. [Online]. Available: https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

[17] United States Department of Energy, "2023 DOE Public Access Plan," United States Department of Energy, 2023. doi: 10.11578/2023DOEPUBLICACCESSPLAN.

[18] National Science Foundation, "NSF Public Access Plan 2.0 Ensuring Open, Immediate and Equitable Access to National Science Foundation Funded Research," NSF 23-104, Feb. 2023. [Online]. Available: https://www.nsf.gov/pubs/2023/nsf23104/nsf23104.pdf

[19] C. Wiley, "Data Sharing and Engineering Faculty: An Analysis of Selected Publications," *Science & Technology Libraries*, vol. 37, no. 4, pp. 409–419, Oct. 2018, doi: 10.1080/0194262X.2018.1516596.