

UC Berkeley

LAUC-B and Library Staff Research

Title

Machine Actionable Plans: Assessing the potential for machine actionable data management and sharing plans (DMSPs) at UC Berkeley

Permalink

<https://escholarship.org/uc/item/6dp5s6b4>

Authors

Foster, Erin

Sackmann, Anna

Publication Date

2026-01-26

Machine Actionable Plans: Assessing the potential for machine actionable data management and sharing plans (DMSPs) at UC Berkeley

Erin Foster
Research Data Management Program Service Lead
edfoster@berkeley.edu

Anna Sackmann
Data Services Librarian
asackmann@berkeley.edu

4 June 2025

Introduction

The primary output of research is often a publication, white paper, or technical publication. However, the data generated and code used for analysis frequently remained unpublished and inaccessible ¹. Over the past decade, researchers, librarians, repository managers, and others involved in the research enterprise have shifted their focus away from publications as the primary output to emphasizing the sharing of underlying data and code ². This shift, which is a crucial piece of Open Science, requires a commitment from researchers to increase transparency in their process.

The Open Science movement has coincided with the introduction of data management and sharing plan (DMSP) requirements from federal and private funders. As a result, institutions must ensure researchers and principal investigators (PIs) comply with agency data-sharing mandates including the August 2022 policy from the Office of Science Technology and Policy (OSTP) on public access for research data and related materials. Even as federal policies evolve, researchers and institutions recognize the value of effective data management and sharing in enhancing research impact and accelerating scientific progress.

The DMPTool has long been a resource for researchers writing data management plans to include in their research grant applications. Managed by the California Digital Library with oversight from an editorial board, the DMPTool is regularly updated to reflect new funding agency requirements. Institutions that utilize the tool typically designate a campus administrator

¹ Jelte M. Wicherts and Marjan Bakker, "Publish (Your Data) or (Let the Data) Perish! Why Not Publish Your Data Too?," *Intelligence* 40, no. 2 (March 1, 2012): 73–76, <https://doi.org/10.1016/j.intell.2012.01.004>.

² Joseph S. Ross and Harlan M. Krumholz, "Ushering in a New Era of Open Science through Data Sharing: The Wall Must Come Down," *Jama* 309, no. 13 (2013): 1355–56.

- often within the institution's library - to provide additional guidance and support to researchers preparing data management and sharing plans.

In 2024, in response to the OSTP memo and the potential to automate the gathering of information from DMPs through machine-actionable processes, the Association of Research Libraries (ARL) and the California Digital Library (CDL) "...partnered to enhance digital research management infrastructures and services" with a focus on enhancing the DMPTool's capabilities for automatically capturing pertinent information through machine actionable DMPs. The project, funded by the Institute of Museum and Library Services (IMLS), includes five pilot institutions:

- Arizona State University
- Northwestern University Feinberg School of Medicine
- Pennsylvania State University
- University of California, Riverside
- University of Colorado, Boulder

And five extended cohort institutions:

- UC Berkeley
- UC San Diego
- UC Santa Barbara
- Stanford
- New York University, Langone

Through their respective projects, these ten institutions are contributing to the development of automated machine-actionable DMSPs that integrate seamlessly with funder and institution data management workflows.

Background

DMSPs are included as part of grant applications and are typically no longer than two pages. The funding agency determines what should be included in these plans³. As mentioned, campus researchers writing plans to fulfill funder requirements have access to updated funder templates through the DMPTool. Researchers can also access the templates for DMSPs - and frequently do - through other methods including downloading the template from an agency website or connecting with colleagues who have previously completed a DMSP for a given agency.

UC Berkeley's Research Data Management (RDM) Program and the Library Data Services Program (LDSP) partner to support the writing and implementation of data management and sharing plans. Campus researchers writing plans to fulfill funder requirements have access to updated funder templates through the DMPTool and guidance from LDSP. The RDM Program assists researchers as they actively work with data through data classification, identifying

³ Hannah Calkins et al., "SPARC Data Sharing Resource Update 2020," August 30, 2022, <https://doi.org/10.17605/OSF.IO/C4PD8>.

appropriate security measures and storage, and connecting storage to sharing for collaboration. Between these two programs, researchers and their data are supported across the lifecycle of a research project.

Machine-actionable Data Management and Sharing Plans (maDMSPs) make the information contained in the plan accessible to not only humans, but machines. Details related to the management and sharing of research data are made available to the multitude of institution stakeholders that are involved in supporting - and planning for - a research project. Allowing for the rich information in these plans to be more broadly available to stakeholders at a given institution can enable better communication, collaboration, and efficiency in a process of conducting research ⁴.

maDMSP pilot project goals

As participants in the extended pilot project cohort, UC Berkeley aimed to enhance the curation and transparency of active research data by connecting researchers with campus supported data compute platforms and storage during the planning stage of the funding process. In the last few years, UC Berkeley and other academic institutions have faced significant storage caps with commercial cloud storage providers. These restrictions, coupled with an increase in data intensive research, have generated an increased urgency to understand and address our research data infrastructure needs, including storage and compute options.

The pilot development of scalable data management infrastructure at UC Berkeley would enable researchers to prospectively review and select available campus platforms for storage and computation. In data management plans, funders already require a description of how data will be analyzed and where it will be stored during the active research phase. By providing researchers with a menu of storage and compute platform options, they will be able to enhance the infrastructure section of the DMP by identifying projected storage allocations and potential funding needs for data storage resources prior to proposal submission.

Additionally, the Research Data Management Program and the UC Berkeley Library hoped to gain insight on DMP content related to researchers' perceived use of storage and computational resources (both non-UCB and UCB supported resources) and the amount of data they aim to generate. And finally, any successful project and program at UC Berkeley required collaboration across campus units. Through this grant, the Berkeley team worked to continue building technical and personnel workflow connections with the university's Sponsored Projects Office grant tool ([Phoebe](#)) to get DMP related information earlier in the process.

The members of the pilot team include:

- Erin Foster, Service Lead, Research Data Management Program
- Anna Sackmann, Data Services Librarian

⁴ Tomasz Miksa et al., "Ten Principles for Machine-Actionable Data Management Plans," ed. Francis Ouellette, *PLOS Computational Biology* 15, no. 3 (March 28, 2019): e1006750, <https://doi.org/10.1371/journal.pcbi.1006750>.

Pilot team members will seek assistance from additional library staff who have provided the RDM Program and LDSP with assistance on previous projects.

Sponsors of this project included:

- Ken Lutz, Director of Research IT and Chief Research Technology Officer
- Salwa Ismail, former Associate University Librarian for Digital Initiatives and Information Technology
- Elizabeth Brashers, Assistant Vice Chancellor & Chief of Staff, Office of the Vice Chancellor for Research
- Kairi Williams, Assistant Vice Chancellor, Research Administration and Compliance

Characteristics of DM(S)Ps included

UC Berkeley's Sponsored Projects Office (SPO) supported this project work by providing team members with access to the institution's database and routing system for grant application proposals. Additionally, SPO provided the team with an extensive list of grant proposals to the NSF and NIH that were successfully funded and completed. This list and access to the database enabled the team to select DMPs for projects that were more likely to have shareable outcomes in the form of data, code, and publications. For the purpose of this project, 10 data management plans were identified from projects that were received over \$250,000 in funding from the following UC Berkeley departments:

- Psychology
- Industrial Engineering and Operations Research
- Chemistry
- School of Education
- California Institute for Quantitative Biosciences (QB3) ([Department of Plant and Microbial Biology](#) and the [Department of Molecular and Cell Biology](#))
- Earth and Planetary Science
- Plant and Microbial Biology
- School of Information
- Electrical Engineering & Computer Sciences

Once identified, DMPs were uploaded into the DMPTool, which connects the plan with other outputs of the grant funded research, including datasets and publications. UC Berkeley's uploaded plans, through the use of persistent identifiers (PIDs) connected to 376 outputs, including peer reviewed articles, book chapters, pre-prints, and datasets. The research output connections are currently being checked and verified by a student library employee. Refining this process and validating the connections, although onerous, will ultimately enable the researcher and institution to record a quantifiable output metric upon completion of the grant.

Findings from DMSPs

A primary outcome of this project was the completion of a qualitative analysis of the 10 DMSPs. In order to determine the potential to be machine-actionable, the team had to assess the contents of existing plans. Using The DART Project rubric as inspiration, five themes were identified that both team members used in the (independent) review and coding of these plans.⁵

Across the 10 plans, a total of 579 tags were applied between the two team members during the qualitative analysis. The table below provides a breakdown of the percentage of tags coded per theme:

Theme	% of tags ⁶
Type of data	21% (n=123)
Data standards	13% (n=78)
Active storage & computation	15% (n=84)
Sharing & reuse	33% (n=189)
Archiving & access	14% (n=79)

Type of data

This theme refers to details related to the type of data generated, collected, or used as part of research. Included as sub-themes are: data collection methods, information about a data provider (if secondary data), and volume of data. The ability to identify how data is being collected or generated, and at what rate or amount, from a DMSP is very helpful information to inform the allocation of appropriate computing and storage support.

All plans included details about the type of data to be collected or generated, often specifying the content as well as if there were multiple types of data involved in the research. Many plans also included some detail as to *how* the data collection was going to occur - both quantitative and qualitative examples of data to be gathered were included in the plans reviewed. Only one plan had a reference to the “volume” of data and that reference was not specific to size or amount, only to some data in the study being “small” and other data being “large”.

Data standards

This theme covers discussion of standards or frameworks used to manage the data (or metadata) involved in the research. Included as sub-themes are: data formats (e.g., file types)

⁵ <https://osf.io/kh2y6/>

⁶ There were additional tags assigned in the coding, which is why the numbers in this table do not total 578 (“Roles & responsibilities” n=12; “Notable quotes” n=14)

and metadata standards (i.e., defined ways of describing and organizing data). Another standard that emerged in reviewing the plans was the inclusion of reference to persistent identifiers (e.g., digital object identifiers or DOIs), which is crucial to allowing for long-term discoverability and sharing of research data.

Data formats (e.g., tiff, PDF, XML) were commonly referenced while metadata standards were less frequently included in the DMSPs. Those plans that did reference metadata standards were either specific about the standard to be used (e.g., MIGS, MIMS) or referred to abiding by the given discipline's (e.g., climate and forecast) expectations when it came to describing and organizing data.

Active storage & computation

This theme covers discussion of how data will be stored and how analysis will occur during the course of research, with an eye towards mention of campus platforms and services. Specific sub-themes included: mention of storage (e.g., Google Drive, Box, hard drives) or computing platform (e.g., VMs, HPC cluster), use of software/programs for analysis, and the development of code and/or scripts.

This information was a particular aspect of the plans that the project team was interested in. In contrast to the other themes, information about a research project's storage and computation is not an explicit section in many DMSP templates. More detail is often required in plans about describing the data to be gathered or collected and how it will be shared (and archived) post-research rather than talking about the "active" parts of working with and storing data. In the case of the DMSPs analyzed, the majority (n=7) made some sort of reference to the storage and/or computing platforms to be used during the research. About half of these mentioned specifically the platform to be used (e.g., UC Berkeley's Google Drive) while the rest kept it vague (e.g., campus storage system).

Sharing & re-use

This theme covers mention of making data and other research outputs (e.g., code, scripts) publicly available and the terms around re-use. This includes any information shared around what data or research outputs will be shared, the timeline for sharing these data and outputs, the repository or platforms for the sharing, and the license (if included) which defines the terms of use around what is being shared.

All plans included a description of what data will be shared, where, and for how long. This is not surprising given the requirement to describe sharing of research data is a fundamental component of all DMSPs. As indicated in the percentage of tags, this theme comprised the majority of tags assigned - in particular because many projects did not just have one output of research. There were multiple outputs mentioned in plans, including: datasets, code and/or scripts, media content, or - in one case - curriculum materials. Roughly six of the plans included reference to a license or terms of use in regards to the outputs to be shared.

Archiving & access

This theme describes how long term retention and access of data and related materials (code, scripts) will occur. Similar to the theme of “sharing & re-use”, this covers mention of the repository or platforms used for the archiving, the timeframe for retention, and if there are any access restrictions (e.g., mediated access).

There was reference to a couple archive platforms amongst the plans - one plan made extensive reference to Merritt Repository, several referred to different disciplinary repositories, and others were less specific in where exactly they would “archive” the research data, but noted they would do so. There were several cases where the use of the word “archive” referred more to long term storage of research outputs rather than archiving in its truest sense.

In most plans, there were no access restrictions in place for the data or other materials generated - the outputs of the research were publicly available. This is in line with the expectations of DMSPs so that is not surprising; however, there were several that spoke to reasons why the data could not be publicly available and included information about how mediated access to the data would be pursued if possible.

Recommendations & next steps

Following the completion of the qualitative analysis of the DMSPs, the next step is to determine how to approach the “machine actionable” aspects of working with DMSPs at UC Berkeley, which will require taking advantage of the enhanced functionality in the DMPTool. However, the DMPTool is not how most of the DMSPs on campus are generated. Given the less-than-ideal engagement in the tool, another step to explore is possible connections to Phoebe (the campus’ grants database) via API to extract the plans included in the submitted grant applications. That would enable the tool to connect plans to research outputs. In many cases though, DMSPs are part of a PDF’d grant application, which requires text extraction. The project team is following up with the developers of Phoebe to learn more and see what particular technical integrations may be able to be established.

Another desire, that was noted in the initial pilot plan, is to identify storage and computation resources mentioned in these plans in order to allocate for those needs. By conducting the qualitative analysis work, the project team gained insight into how these types of resources were (or were not) referred to in the sample of plans analyzed. This background work provides a solid foundation for considering how, realistically, a storage and computation resource allocation might occur. One of the primary cohort sites, UC Riverside, produced an “allocation dashboard” as an outcome from their project work. One potential next step is to see if that dashboard, combined with the understanding of how these resources are referred to in plans, could be utilized to estimate how much of a given storage and/or computing resource a researcher might need when completing their research described in a DMSP. That kind of insight would be helpful for, not just the researchers, but the staff and units that support and run the systems and platforms that will be used. In an ideal scenario, it could allow for more mindful provisioning of resources, as well as provide information to help with budgeting and future planning on behalf of

the institution. It is unlikely that the level of detail provided in these plans - regarding storage and computation needs - is enough to allow for such an ideal outcome, but it is worth exploring how these plans can help with more upfront thinking around resources to be used.

Bibliography

Calkins, Hannah, Patricia Condon, Jonathan Petters, Rachel Woodbrook, and Reid Boehm.

“SPARC Data Sharing Resource Update 2020,” August 30, 2022.

<https://doi.org/10.17605/OSF.IO/C4PD8>.

Miksa, Tomasz, Stephanie Simms, Daniel Mietchen, and Sarah Jones. “Ten Principles for Machine-Actionable Data Management Plans.” Edited by Francis Ouellette. *PLOS Computational Biology* 15, no. 3 (March 28, 2019): e1006750.

<https://doi.org/10.1371/journal.pcbi.1006750>.

Ross, Joseph S., and Harlan M. Krumholz. “Ushering in a New Era of Open Science through Data Sharing: The Wall Must Come Down.” *Jama* 309, no. 13 (2013): 1355–56.

Wicherts, Jelte M., and Marjan Bakker. “Publish (Your Data) or (Let the Data) Perish! Why Not Publish Your Data Too?” *Intelligence* 40, no. 2 (March 1, 2012): 73–76.

<https://doi.org/10.1016/j.intell.2012.01.004>.

Appendix 1: [Codebook](#)

Appendix 2: [Export of tags](#)

Appendix 3: [Notes on plan coding](#)