University of Colorado Boulder - Machine Actionable Plans (MAP) Pilot Project Report

Executive Summary

This report details the activities, lessons learned, and recommendations from the University of Colorado Boulder (CU Boulder) pilot project as part of the Machine Actionable Plans (MAP) Pilot led by the California Digital Library (CDL) and the Association of Research Libraries (ARL) to enhance infrastructure and services around machine actionable Data Management and Sharing Plans (maDMSPs). CU Boulder was one of five main pilot sites competitively selected to participate in this project, and the project team represented a cross-campus collaborative group led by the Center for Research Data & Digital Scholarship (CRDDS) with representation from Libraries, Research Computing, the Research & Innovation Office (RIO), the Faculty Information System (FIS), and the Laboratory for Atmospheric and Space Physics (LASP).

Based on the pilot project activities, this report makes the following recommendations in order to leverage maDMSPs to enhance the value of the research data produced by CU Boulder researchers:

- **Enhance collaboration across campus with regard to public access to research data and DMSPs.**
- **Ensure research data is integrated into a holistic strategy for research information management across campus.**
- **Develop a coordinated strategy and approach for persistent identifiers (PIDs) at the campus level.**

These recommendations are explained in further detail at the end of the report following a complete description of the pilot project objectives, activities, lessons learned, and next steps.

Background

In late 2023, the University of Colorado Boulder (CU Boulder) was selected as one of five main pilot sites for a project led by the California Digital Library (CDL) and the Association of Research Libraries (ARL) to enhance research management infrastructure and services around machine actionable Data Management and Sharing Plans (maDMSPs). An additional five institutions were selected for the maDMSP extended cohort that engaged closely with the main pilot cohort. While DMSPs are typically thought of as requirements researchers must meet as part of grant proposals to federal funding agencies, maDMSPs offer the potential to enhance the value of research data produced at institutions by making data more discoverable and connected to other parts of the research ecosystem. In recent years, maDMSPs have emerged as key mechanisms for federal funding agency plans for public access to research data; however, this landscape is rapidly changing as public access policies at many agencies are still being released.

Funded by the Institute of Museum and Library Services (IMLS), award LG-254861-OLS-23, this project focused on advancing capabilities of maDMSPs by enhancing the DMPTool (an online customizable tool for creating maDMSPs) and by exploring how maDMSPs can be integrated across the existing research infrastructure at the pilot institutions to automate and improve data management processes. The project supported strategic planning for libraries and other campus units as they implement new policies, workflows, and technical solutions to build local capacity for effective data management and local research coordination. The pilot institutions helped shape the development of maDMSP functionality in a widely-used tool for creating DMSPs (DMPTool) and gained valuable early experience with new approaches to enable more automated and connected research data management. The full list of pilot institutions included:

- Arizona State University (main cohort)
- Northwestern University Feinberg School of Medicine (main cohort)
- Pennsylvania State University (main cohort)
- University of California, Riverside (main cohort)
- University of Colorado Boulder (main cohort)
- New York University Langone Health (extended cohort)
- Stanford University (extended cohort)
- University of California, Berkeley (extended cohort)
- University of California, San Diego (extended cohort)
- University of California, Santa Barbara (extended cohort)

The CU Boulder project team represented a cross-campus collaborative group led by the Center for Research Data & Digital Scholarship (CRDDS). The project team included representation from Libraries, Research Computing (in the Office of Information Technology (OIT)), the

Research & Innovation Office (RIO), the Faculty Information System (FIS) in OIT, and the Laboratory for Atmospheric and Space Physics (LASP). Project team members included:

- Andrew Johnson, Head of Data and Scholarly Communication Services, Libraries/CRDDS (project lead)
- Thea Lindquist, Professor and Executive Director, Libraries/CRDDS
- Matthew Murray, Data Librarian, Libraries/CRDDS
- Adi Ranganath, Data Librarian, Libraries/CRDDS
- Layla Freeborn, Associate Director of User Services, Research Computing
- Shelley Knuth, Assistant Vice Chancellor for Research Computing, OIT
- Barb Schnell, Associate Director of Secure Research and Computing, Research Computing, OIT
- Samuel Oskar Klopsch, NCAR/UCAR Developer in Residence, Libraries
- Vida Sabeti, Digital Library Senior Software Developer, Libraries
- Jamie Wittenberg, Assistant Dean for Research & Innovation Strategies, Libraries
- Doug Lindholm, Data Systems Software Engineer and Open Science Co-chair, LASP
- Karen Regan, Associate Vice Chancellor for Research & Innovation, Research Development, RIO
- Don Elsborg, Lead Architect, FIS, OIT (retired)
- Alex Viggio, Director, FIS, OIT

Project Objectives

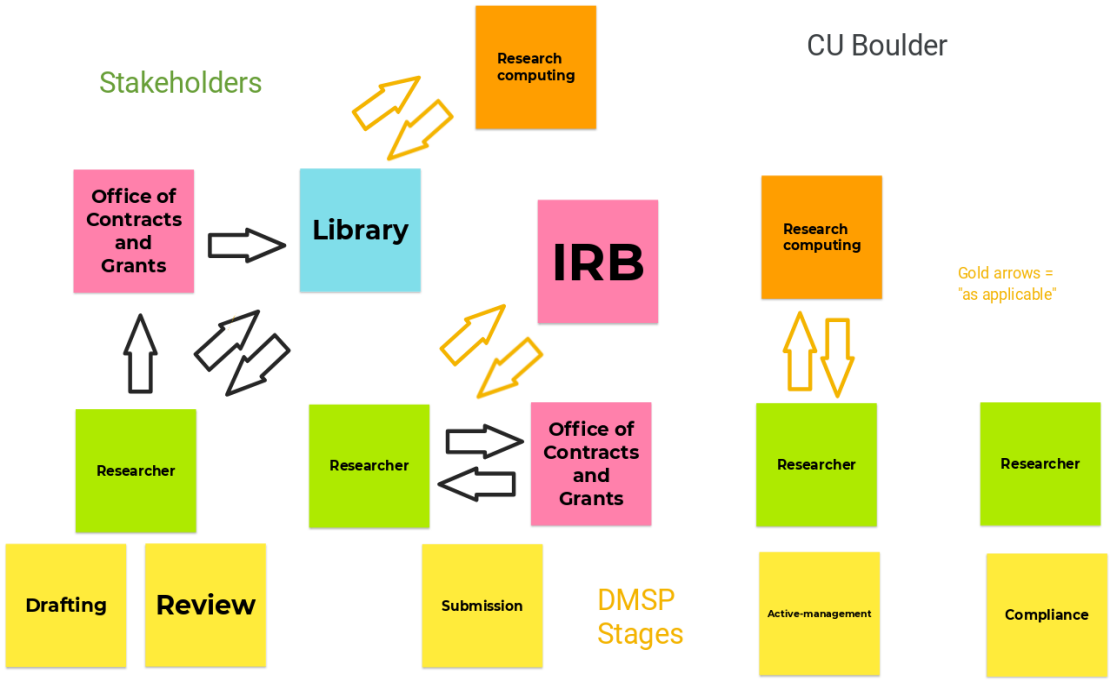
CU Boulder's maDMSP pilot project initially cast a wide net in identifying potential use cases and stakeholders across campus. While we ultimately focused more of our efforts on some of the use cases than on others, the initial broad scope was valuable in allowing us to have multiple directions to pursue as roadblocks emerged for some of our objectives. This scope was also valuable in raising awareness of maDMSPs across campus even if the immediate work of the pilot ended up involving some stakeholders more directly than others. Our initial scope sought to develop workflows and automated mechanisms to connect and communicate across campus units involved with research data management and public access. This included exploring integrations (e.g., via APIs) with campus systems, including the CU Scholar institutional repository, the CU Boulder Elements (CUBE) research information management system, the CU Experts researcher profile system, and the Laboratory for Atmospheric and Space Physics (LASP) scientific dataset metadata repository. In addition, the project aimed to develop sociotechnical communication workflows to alert units that provide support, services, and infrastructure for research data as it moves through the lifecycle of a project from the pre/post-award stages to data collection and analysis to providing public access and preservation in the final phase. The units that found potential value in such alerts included CRDDS, Research Computing, Libraries, RIO, FIS, and LASP. The following are all of the specific objectives we intended to accomplish during the pilot project:

- Enable better planning for enhancement and scaling of infrastructure and services by automating alerts for personnel when specific units or the infrastructure they support (e.g., institutional repository, large-scale storage system, HPC infrastructure, cloud computing resources) are included in DMSPs from awarded proposals.
- Improve support for researchers using campus cyberinfrastructure (or who could benefit from such resources) for machine learning and artificial intelligence by automating alerts for personnel when DMSPs include key terms related to these topics.
- Streamline and centralize collection of information about published datasets related to awarded DMSPs for reuse in other systems (e.g., annual reporting for campus and funders, researcher profile systems, etc.) by creating an automated pipeline of metadata about datasets deposited in the CU Scholar repository into maDMSPs from related awarded proposals, and by evaluating the viability of a similar process for datasets deposited in other repositories (e.g., disciplinary repositories).
- Support researchers with compliance needs across campus by developing capabilities for tracking data classification as it moves through the lifecycle of a project (e.g., when it is collected or received, where it is stored, how long it needs to be retained, when it can/must be deleted, etc.), and creating alerts for various campus entities involved with regulated data (Institutional Review Board (IRB), Office of Contracts and Grants (OCG), OIT Security, Research Computing, etc.).
- Track LASP's scientific data deliverables to national data repositories for both small research grants and large spacecraft missions by linking dataset metadata to the planned data management activities as recorded in maDMSPs, which will extend the knowledge graph representing the provenance of the datasets.

Project Activities

The pilot project activities began in early 2024 and ran through early 2025. These activities included a number of local efforts aimed at achieving our CU Boulder project goals as well as regular meetings with both the ARL/CDL project leads and the entire project cohort of pilot institutions. As part of this work with the entire cohort, we documented the existing DMSP workflows at our institution (see Figure 1), and we contributed to testing new maDMSP functionalities in the DMPTool. In October of 2024, we hosted the ARL/CDL project leads for a site visit, which included meetings with our CU Boulder project team, interviews with campus stakeholders, and two public presentations about our pilot project at CRDDS and LASP, respectively.

Figure 1. Current DMSP workflows at CU Boulder.



In the remainder of this section, we will describe the local activities we undertook to meet our CU Boulder-specific project goals. Early in the project, we encountered several barriers to some of our proposed objectives. First, in working with our project team partners from RIO, we discovered that it was more difficult and labor-intensive than expected to obtain a large enough sample of pre/post-award DMSPs for effective testing of many of our alert-based objectives, especially due to the lack of detail in existing DMSPs. We also encountered concerns related to the data classification tracking objective as we discovered that requirements for the security of any systems used to track such data likely needing to adhere to the same security standards as the systems for storing the actual data. In response to these barriers, we directed early pilot project efforts toward the "Streamline and centralize collection of information about published datasets…" and "Track LASP's scientific data deliverables…" objectives, both of which focused on post-award DMSPs. This allowed us to work with existing systems that the project team was more familiar with (e.g., CUBE) and/or that were more directly under the control of the project team (e.g., CU Scholar, LASP data systems). This shift also maintained the involvement of LASP, which was our key NASA-funded use case (NASA being one of the primary research funders for CU Boulder) as well as our research institute use case. To work toward these two objectives, we engaged in the main activities described below.

First, we identified all of the current CU Boulder systems that already interact with DMSPs and the DMPTool as well as the internal and external data sources we currently use or could use to enhance the machine-actionability and completeness of post-award DMSPs with regard to other parts of the research ecosystem (e.g., researchers, grants, publications, published datasets) (see

Figure 2). We then identified where improved integration (either technical or sociotechnical) among these systems could help us identify where we need to focus our efforts to meet our objectives (see Figure 3). Figure 4 provides additional descriptions of the internal data sources used in both current and desired state scenarios. In addition to these activities aimed at identifying connections among data sources to prioritize at the campus level, we mapped how maDMSPs might integrate into the infrastructure at LASP as a key use case for individual units (in this case a research institute), which often have their own domain-specific data infrastructure (see Figure 5). It is important to note that persistent identifiers (PIDs) are essential to all current and desired integrations in these diagrams. PIDs currently in use in the data sources we identified and worked with during the pilot project include DataCite DOIs (datasets), CrossRef DOIs (publications), ORCIDs (individual researchers), RORs (funders/institutions), DMP IDs (maDMSPs), and more. These PIDs help to accurately identify and track all of the entities in the research ecosystem, and the metadata related to these PIDs can better enable automated connections among these entities in maDMSPs and other systems.

Figure 2. Current state of maDMSP-related information at CU Boulder.

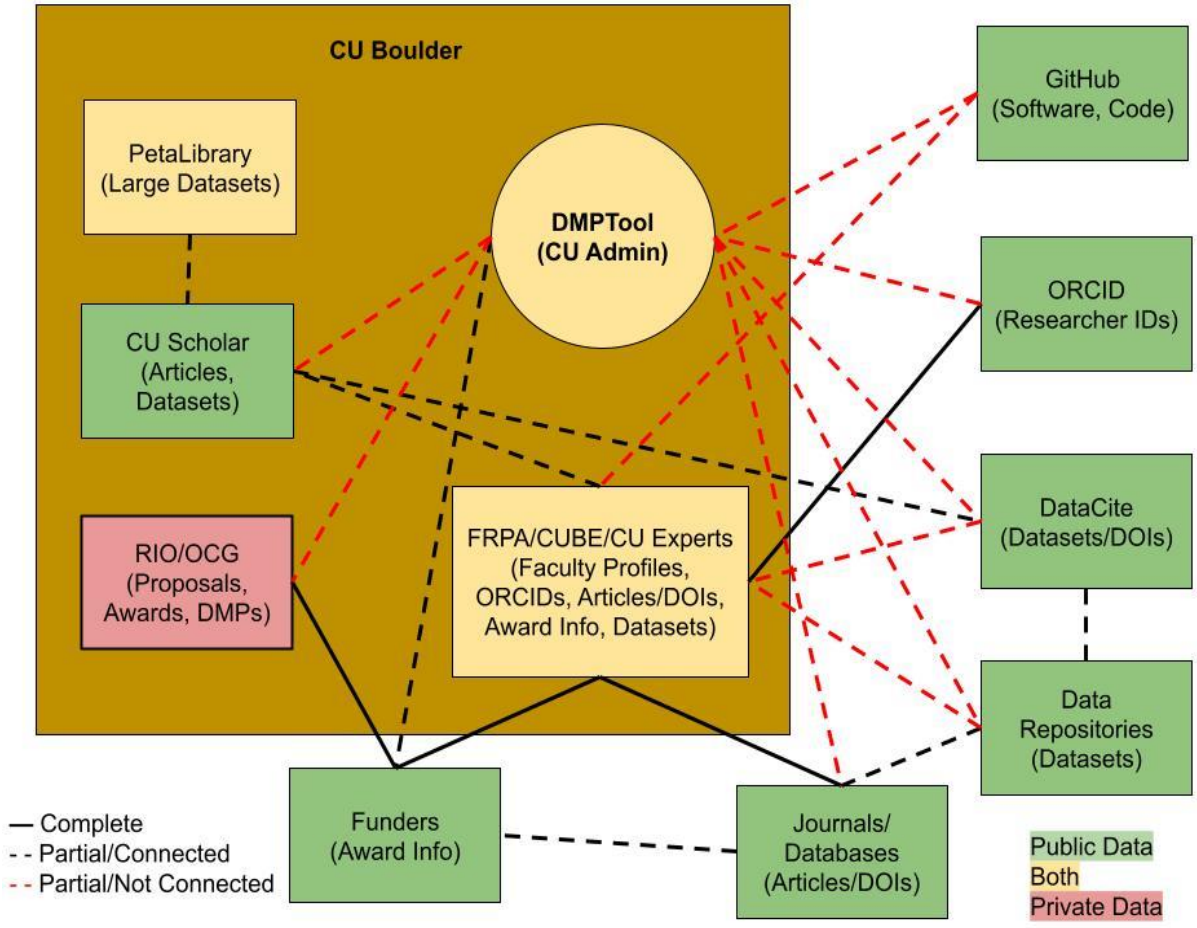


Figure 3. Desired state of maDMSP-related information at CU Boulder.

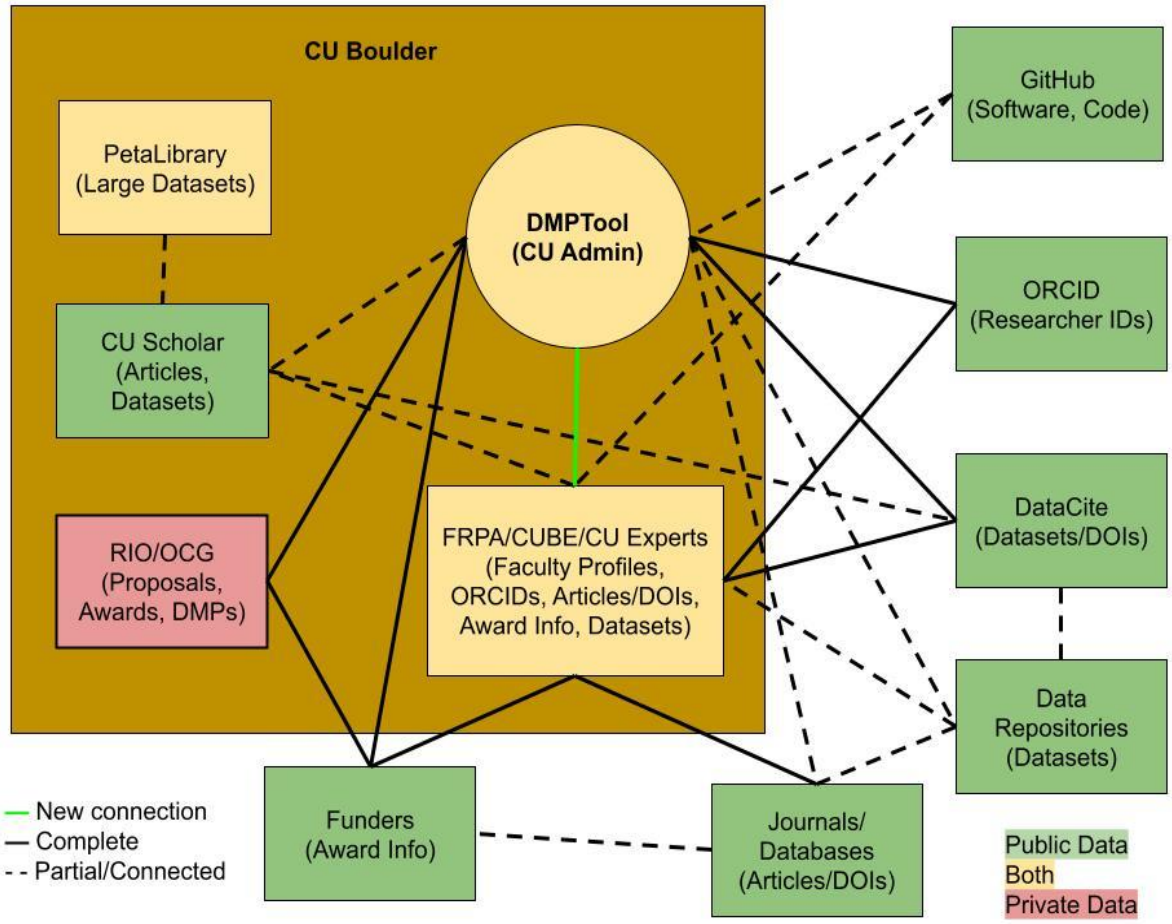
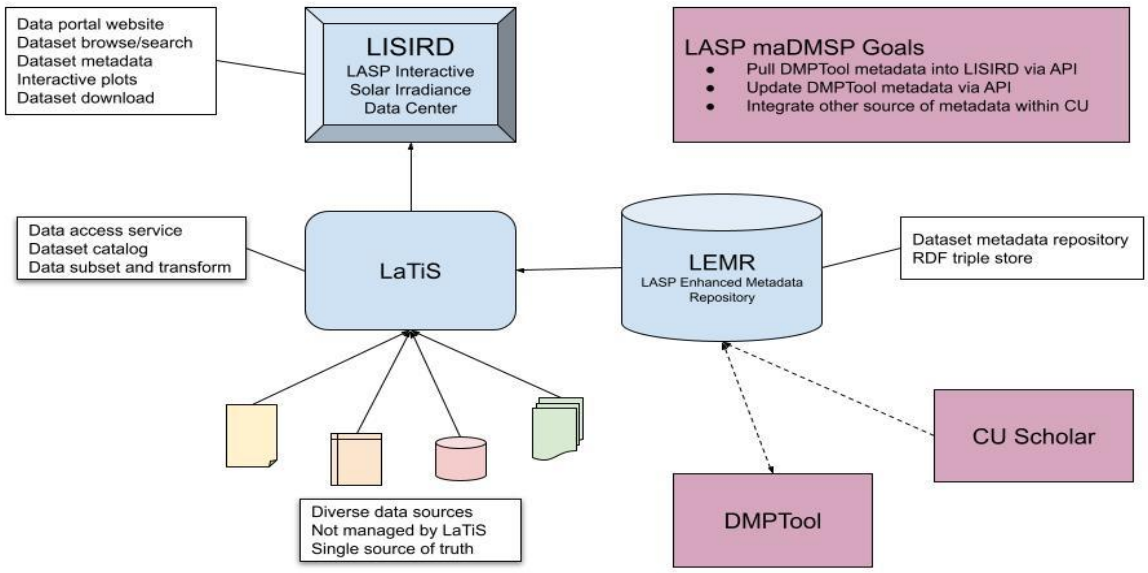


Figure 4. Descriptions of internal CU Boulder data sources used in Figures 2 and 3.

CU Scholar (Articles, Datasets)	University of Colorado Boulder's open access institutional repository
PetaLibrary (Large Datasets)	University of Colorado Boulder's Research Computing service that supports the storage, archival, and sharing of "big" data (e.g., TBs of storage)
RIO/OCG (Proposals, Awards, DMPs)	RIO: Research & Innovation Office OCG: Office of Contracts and Grants
FRPA/CUBE/CU Experts (Faculty Profiles, ORCIDs, Articles/DOIs, Award Info, Datasets)	FRPA: Faculty Report of Professional Activities CUBE: CU Boulder's instance of Symplectic Elements CU Experts: Public faculty profiles (uses CUBE data)

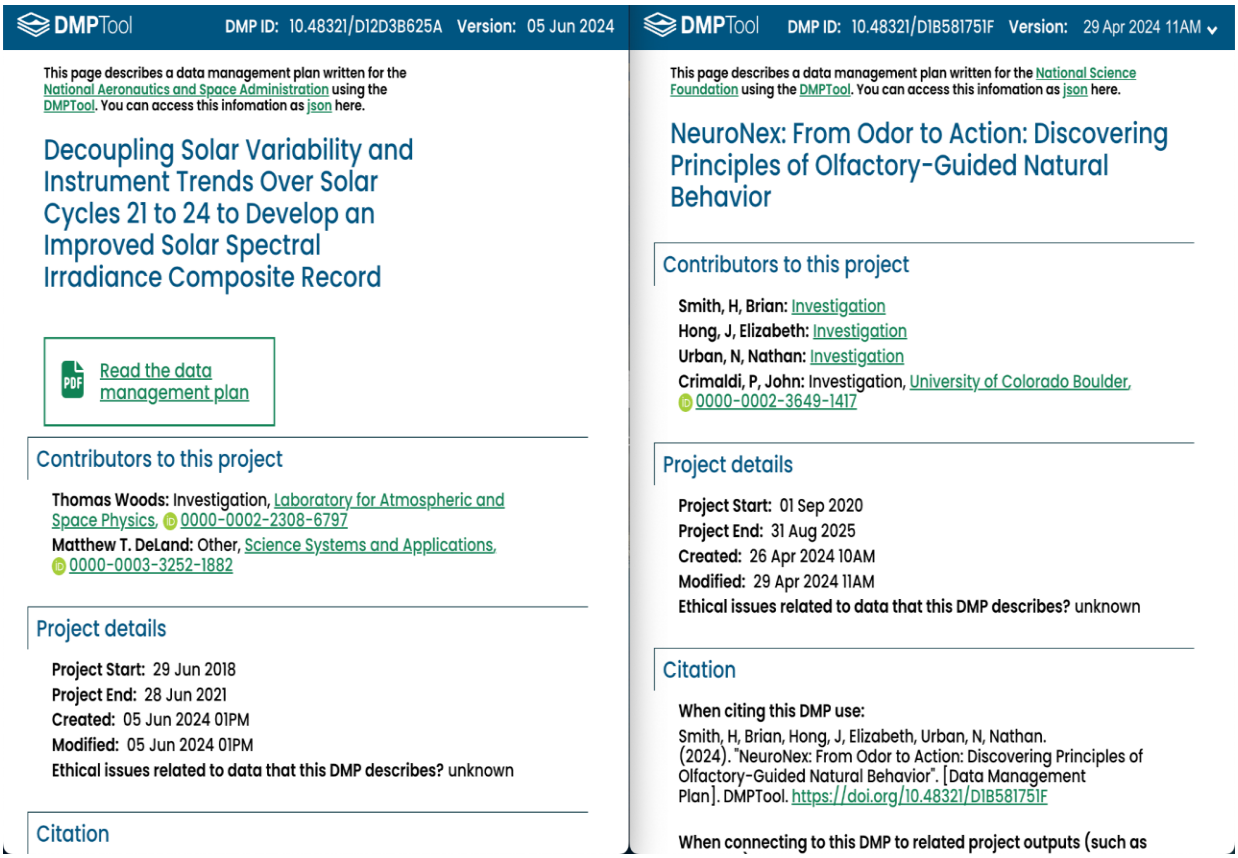
Figure 5. LASP data infrastructure with desired connections to maDMSPs and campus systems.



Following these diagramming activities with campus and LASP infrastructure, we identified two case study maDMSPs from large research projects to use to test what information we could identify using existing data sources to populate actual maDMSPs (even if this needed to be done manually). We chose maDMSPs from the NSF-funded "NeuroNex: From Odor to Action: Discovering Principles of Olfactory-Guided Natural Behavior" project as well as the NASA-funded "Decoupling Solar Variability and Instrument Trends Over Solar Cycles 21 to 24 to Develop an Improved Solar Spectral Irradiance Composite Record" (see Figure 6). These grants were chosen due to the familiarity individual project team members had with them, which would allow us to better evaluate the completeness of data sources in identifying types of information related to the grant (e.g., publications). We then searched each of the data sources listed in the campus-level diagram (see Figure 2) for information related to these two use cases, and we documented what information we found from each data source. During this process, it became apparent that identifying published datasets (which can also include software/code) associated with specific grants is incredibly difficult even for grants with which project team members were familiar. We had more success with identifying publications (e.g., journal articles) related to these grants, but we also found quite a bit of variability in the number and accuracy of results for publications associated with specific grants across the available data sources we evaluated (e.g., CUBE, Dimensions, NSF Award Search, DMPTool, etc.). As expected, ORCID and sources that integrate with it (e.g., CUBE/CU Experts) were reliable sources of information about individual researchers for those who have an ORCID and keep it updated with publicly available

information; however, it seems that ORCID integration with CU Experts has not been possible since 2022.

Figure 6. Screenshot of maDMSPs for two CU Boulder case studies.



Finally, we began initial testing to understand how to use the DMPTool API, what it can be used for, and what use cases we might have for it from our various stakeholder groups. Particularly, we are interested in using the API for the enhancement of post-award maDMSPs to potentially automate any of the more manual activities described above, and we are still planning to investigate this possible use case.

In addition to the other pilot project activities, we submitted four conference proposals related to aspects of our maDMSP pilot in order to share what we learned from our project with large national audiences, which allowed us both to showcase CU Boulder as a leader in this area and to extend the insights we gained to other interested institutions. We submitted proposals to the following conferences: Association of College and Research Libraries (ACRL) 2025 Conference, Research Data Access and Preservation (RDAP) Summit 2025, Coalition for Networked Information (CNI) Spring 2025 Membership Meeting, and Open Repositories 2025. Three of these (ACRL, CNI, and RDAP) were in collaboration with the project leads and/or the other pilot institutions, and one proposal (Open Repositories) was submitted solely by members

of our CU Boulder pilot project team. While the ACRL proposal was not accepted, CU Boulder project team members successfully presented at RDAP, CNI, and Open Repositories. CU Boulder project team members also contributed to Pennsylvania State University's "Machine-Actionable Data Management and Sharing Plan Workshop" and two of the four sessions in the culminating public webinar series "Insights from the Machine Actionable Plans (MAP) Pilot" hosted by ARL and CDL.

Successes and Lessons Learned

We had significant success in developing a deeper understanding of the campus environment with regard to DMSPs and where they fit into the larger research information ecosystem. Relatedly, we were successful in raising awareness of maDMSPs as a concept and a potential source of valuable information that could support the work of a variety of campus stakeholders, including all the units involved with the pilot team (CRDDS, Libraries, Research Computing, RIO, FIS, and LASP). While we perhaps did not get as far with developing and testing automated methods of enhancing post-award maDMSPs (as described above), we laid the groundwork for doing so in the future by identifying which data sources we would use for such workflows. With regard to specific data sources, we found that we have a variety of options for sources of publications related to grants, but despite identifying DataCite as a logical source of information about published datasets, we found that those datasets were very difficult to associate with funded awards. Directly connecting datasets with grants remains a significant challenge.

We also began identifying where information from maDMSPs could be useful to pull into other campus systems, but we focused more effort on the other direction of the flow of information during the pilot (i.e., campus/external sources feeding into maDMSPs). As described above, we did run into significant barriers to some of our initial alert-based objectives; however, we still gained valuable lessons about where we should direct our energy in order to have a better chance of success and adoption of maDMSPs going forward.

The pilot project activities were also helpful in identifying several areas where it would be useful for us to update and add to our existing institution-specific guidance in the DMPTool for CU Boulder users who log in with their institutional credentials. This included adding questions to DMSP templates about planned use of specific CU Boulder resources (e.g., CU Scholar repository, PetaLibrary large-scale data storage). Previously, we had included suggested language in our customized DMPTool guidance for people to copy/paste into template sections if they intended to use these resources, but we hope that adding specific questions about these resources to templates might encourage more researchers to include them in DMSPs when applicable. This could also make it easier to track mentions of these resources in DMSPs than by searching the full text. We also identified the need to update our DMPTool guidance regarding sensitive/secure data.

Identifying as many stakeholders as possible in the early stages of the project worked out well in surfacing potential use cases that we might not have thought of otherwise. This also helped us discover and confirm roadblocks to some of our objectives (e.g., being able to work with RIO to see what the actual process of obtaining pre/post-award DMSPs would look like). It was also incredibly valuable to have LASP involved as a domain-specific research unit since their involvement provided key insights into what infrastructure looks like in individual data-intensive units as well as the perspectives of data professionals and researchers working closely with active research data. LASP's involvement also brought in-depth knowledge of additional funders (e.g., NASA) and types of research (e.g., data related to space missions) to the project.

As we heard from other institutions involved in the pilot, getting access to all post-award DMSPs and/or getting included in the pre-award pipeline for DMSPs would be a significant challenge for us. Like other institutions, only a relatively small number of researchers at CU Boulder currently use the DMPTool to generate DMSPs for their grant proposals. Similarly, while we have developed a strong referral pathway from our Office of Contracts and Grants to CRDDS for assistance with creating and reviewing DMSPs, the number of researchers who actually reach out to CRDDS is only a subset of the total number who submit proposals that require DMSPs. With those being the two main points of pre-award contact where we could try to implement new maDMSP workflows, we learned from the outset that we would only be reaching a small number of grant proposals that way. For post-award DMSPs, we are currently required to get permission from each PI to access their DMSPs, so this would also be a very labor-intensive process if we were to develop a workflow designed to reach many/all post-award DMSPs at our institution. Similarly, we confirmed during the pilot that the researchers who do use the DMPTool are very unlikely to return to their DMSPs after the proposal stage (even just to update whether the proposal was awarded or even submitted). This all led us to the conclusion that the most effective intervention point for us at this stage would be to focus on post-award DMSPs from projects we already work with closely where we could show value by creating an important resource for researchers or grant teams that would require as little input as possible from the researchers themselves. The hope is that we could then use maDMSPs from these projects as exemplars to show to other researchers and stakeholders on campus to promote wider adoption of maDMSPs. Identifying and being able to articulate the value of these maDMSPs to researchers and the institution will be key to achieving this goal.

Next Steps and Recommendations

In terms of next steps beyond the pilot phase, we are continuing to test and evaluate use of the DMPTool API as described above. We are also closely monitoring changes at federal funding agencies that could impact how we implement maDMSPs. As noted, we are planning to make updates to the local guidance and templates we provide to researchers via the DMPTool that we identified during the pilot, but the timing of this might also be impacted by changes at the federal

level. Since one of the key gaps we found in our ability to integrate datasets with the wider research ecosystem was the lack of direct connection between datasets and associated grants, we began updating our local practices with creating DataCite metadata for the Digital Object Identifiers (DOIs) we register for all datasets in the CU Scholar repository. For at least this subset of datasets produced at CU Boulder, these changes to our local practices could help to improve the metadata about related grants in DataCite, which was the primary source we identified for information about published research data. We are also planning to continue to finalize and promote our exemplar maDMSPs to demonstrate potential use cases and impact of maDMSPs both in general and in regard to campus conversations around research information management strategy. Finally, many of our pilot project team members have indicated a desire to keep in touch and continue to build off the progress we have made so far during the pilot, so we intend to keep those conversations going to see where future collaboration could be valuable.

Based on all the pilot project activities, lessons learned, and next steps identified, we recommend the following in order to leverage maDMSPs to enhance the value of the research data produced by CU Boulder researchers:

- **Enhance collaboration across campus with regard to public access to research data and DMSPs.** It will be essential to monitor how policies around public access to research data at the federal level continue to evolve, but assuming that these policies remain priorities for funders, it will be beneficial to evaluate current workflows and explore possible deeper integrations across campus units that would allow all researchers and interested stakeholders to take advantage of maDMSP functionality both pre- and post-award. This could result in more competitive grant proposals, more streamlined compliance efforts, and greater impact of publicly available research data.
- **Ensure research data is integrated into a holistic strategy for research information management across campus.** Research data remains one of the most difficult entities to track and connect with the rest of the research ecosystem. As this pilot project demonstrated, research data is managed and published using unit-level, campus-level, and external infrastructure and repositories. In order to realize the full potential of research data to demonstrate impact of research beyond traditional outputs like publications, it is necessary to ensure that sociotechnical infrastructure is in place to better track and associate datasets with the other research information sources in use across campus and externally. Ideally, integrations across these sources, including maDMSPs, should be as automated as possible.
- **Develop a coordinated strategy and approach for persistent identifiers (PIDs) at the campus level.** PIDs currently in use in the data sources we identified and worked with during the pilot project include DataCite DOIs (datasets), CrossRef DOIs (publications), ORCIDs (individual researchers), RORs (funders/institutions), DMP IDs (maDMSPs), and more. These PIDs help to accurately identify and track all of the entities in the

research ecosystem, and the metadata related to these PIDs can better enable automated connections among these entities in maDMSPs and other systems. Adoption and use of PIDs can benefit greatly from institutional coordination in helping to ensure all research entities associated with CU Boulder have PIDs and that stakeholders across campus know how to find and utilize these PIDs. In turn, such efforts will allow maDMSPs (and other systems) to more easily leverage PIDs to automate connections among datasets, publications, grants, individual researchers, etc. Such automated connections will provide a more comprehensive picture of the full impact of CU Boulder research.