2021 Sautter Award Submission: DMPHub

Submitter
Marisa Strong
Application Development Manager, California Digital Library, UCOP
https://orcid.org/0000-0002-4229-8939

Project Team

Maria Praetzellis
Product Manager
California Digital Library, UCOP
https://orcid.org/0000-0001-5047-3090

Brian Riley
Technology Lead
California Digital Library, UCOP
https://orcid.org/0000-0001-9870-5

Project Summary

As a National Science Foundation (NSF) grant funded project, the DMPHub was to transform data management plans (DMPs) from their static PDF form into a dynamic and machine-actionable hub to document and disseminate the products of research activity.
Use cases included:

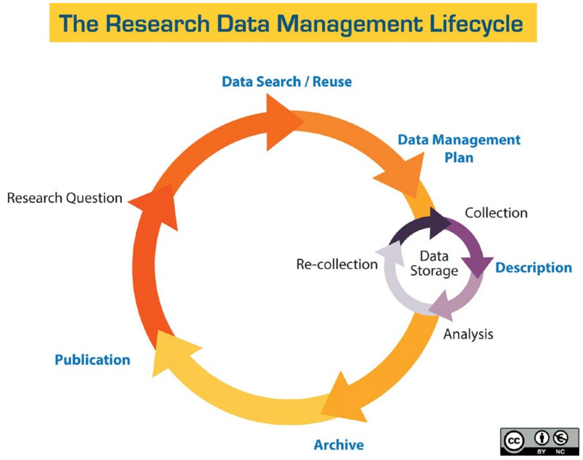
- implementing a set of common standards and exchange protocols for DMPs to enable information to flow between DMPs and existing research information systems;
- leveraging persistent identifiers to trigger notifications across systems enabling stakeholders to plan resources, connect research outputs, automate reporting and monitoring, receive credit and promote data discoverability, reuse and reproducibility;
- building a repository of metadata for the networked DMPs;
- document in a central space displaying connections, outputs, and activities of a DMP

Problem Statement and Goal

Open Data is the movement to make research datasets open, thereby enabling the sharing, reuse, and transparency of research findings. It is critical for advancing science and humanities research in the digital age. Open data policies are proliferating worldwide and researchers are now required to submit Data Management Plans (DMPs) with most grant proposals that describe the data they will produce and plans for sharing and preserving it. When written effectively, DMPs clarify how researchers will effectively disseminate and share research results, data, and associated materials. Researchers do not always know exactly what data they will produce at the beginning of a project, however. Furthermore, they have no incentives or easy methods for updating a DMP to keep things organized over the course of their research, which can lead to poor data practices and chaotic, unusable data shared at the end. DMPs in their current, static document form pose similar challenges for other stakeholders across the research ecosystem, e.g., funders who must monitor compliance manually. The DMPHub project aims to solve this problem by providing a new platform repositioning DMPs as hubs of the networked research ecosystem, facilitating and advancing the research process for all stakeholders.

Brief History of Data Management Plans

The DMPTool preceded the DMPHub and is a free, open-source, online application in use since 2011. The tool was developed in direct response to demands from funding agencies, such as the National Science Foundation (NSF) and the National Institutes of Health (NIH), that researchers plan for managing their research data. Today, funding agencies still require that a researcher outline their data management plan as part of their proposals and the DMPTool continues to fulfill that need. Today's DMPs describe data that will be acquired or produced during research; how the data will be managed, described, and stored, what standards are used, and how data will be handled and protected during and after the completion of the project. The importance of the DMP however has grown beyond that original scope and the need to reference those plans at later stages of a research data's life cycle has emerged.



Research Data Management Lifecycle example from UC Santa Cruz Library

Evolution of the DMPHub

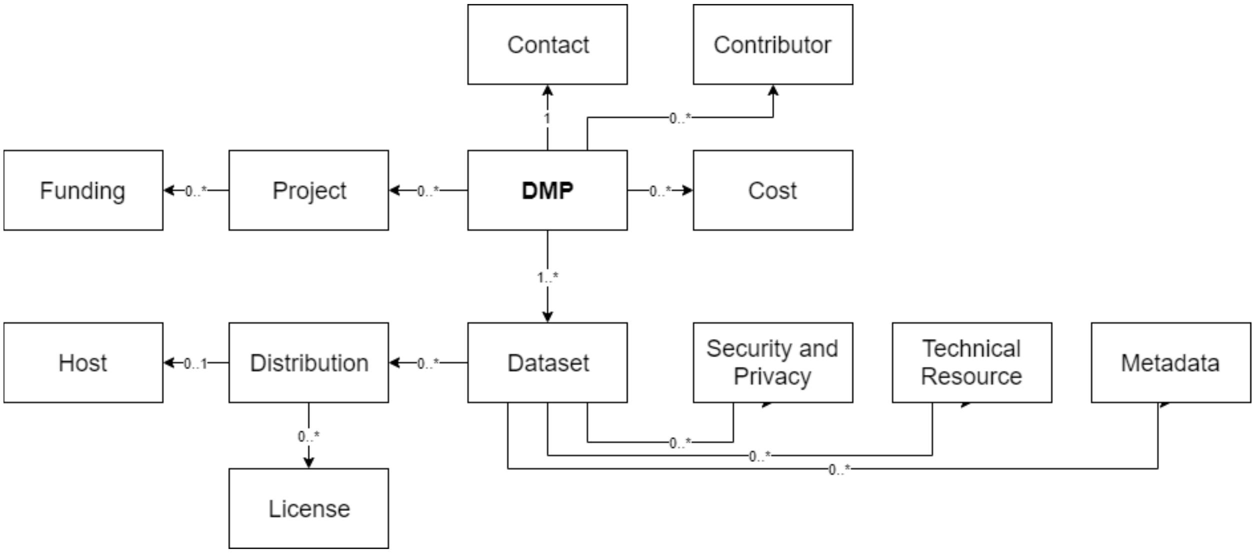
The DMPHub Project evolved over a set of phases. The first phase began with the award of an NFS EAGER grant to prototype and develop a proof of concept. In phase two, the team implemented the proof of concept within the DMPTool ecosystem and phase three extended this work into partnerships and integrations within UC research centers and other research projects across the globe. Throughout all phases, collaborations with key stakeholders and implementers occurred to support the project. They included DataCite, a leading global non-profit organisation that provides persistent identifiers (Digital Object Identifiers - DOIs) for research data and other research outputs; Research Data Alliance (RDA), a community-driven initiative whose mission is to build the social and technical bridges which enable open sharing and re-use of data; Biological and Chemical Oceanography Data Management Office (BCO-DMO); University of California Natural Reserve System, a network of 41 university-run field stations; and the Digital Curation Centre (DCC), a world-leading centre supporting active management of data through the research lifecycle.

Proof of Concept

Persistent identifiers (PIDs) are globally unique labels used to name an object and are guaranteed to be managed and kept up to date over a defined time period. Resolving an identifier gets information unique to that thing, its' metadata, along with metadata changes made throughout its lifecycle.

The proof of concept aimed to investigate if connections could be made between a DMP and the research project's subsequent outputs. This would demonstrate the potential value of a more structured format (metadata) and a persistent identifier to describe and access a DMP. Together the DMP metadata and identifier would result in a machine-actionable DMP providing the ability of DMPs to be findable, accessible, interoperable and reusable (FAIR) with none or minimal human intervention. This automated information exchange transforms DMPs from static narrative documents into structured, interoperable data fed across stakeholders; linking metadata, repositories, and institutions, and allowing for notifications, verification, and reporting in real-time.

Working in partnership with RDA, we defined and became early adopters of a new metadata standard for DMPs, helping to shape future standards for data management planning. This RDA DMP Common Standard expresses information from traditional DMPs in a machine-actionable way, providing interoperability between systems representing information across the DMP lifecycle. The DMPTool API was updated, in partnership with DCC, to export DMPs as JSON defined by the RDA common standard. New metadata values were included in the newly released DataCite Schema 4.4 to support creation of a DMP identifier: DMP ID.

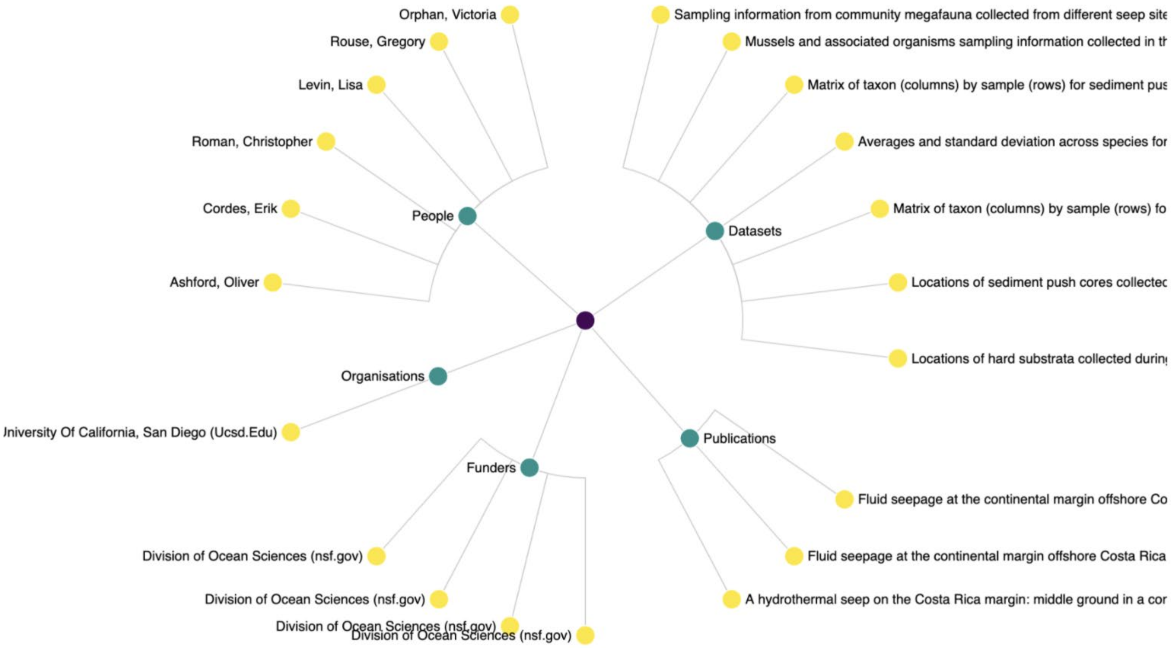


Structure of a machine-actionable DMP defined by the RDA DMP Common Standard

Partnering with BCO-DMO and their data repository for biological research projects funded by the NSF, we acquired historical DMPs that had known connections with published articles, software, and datasets.

These static PDF DMPs were converted into the new metadata format and stored within the prototype DMPHub system and assigned DMP IDs. With this DMP ID and the DMPHub's API, queries could be made to discover and connect other identifiers related to the DMP.

Our new DMP metadata records now had their own DMP ID as well as a host of other related identifiers (e.g. people with ORCIDs, research organizations with RORs, funders with Crossref Funder Ids, articles, publications, and datasets with DOIs.) that could be used to visualize connections with organizations, researchers and published works. Collaborating with DataCite, we were able to explore those connections through their PID graph system.


Orphan, Victoria
Rouse, Gregory
Levin, Lisa
Roman, Christopher
Cordes, Erik
People
Ashford, Oliver
Sampling information from community megafauna collected from different seep site
Mussels and associated organisms sampling information collected in th
Matrix of taxon (columns) by sample (rows) for sediment pus
Averages and standard deviation across species for
Matrix of taxon (columns) by sample (rows) fo
Datasets
Locations of sediment push cores collectec
Locations of hard substrata collected durin
Organisations
Jniversity Of California, San Diego (Ucsd.Edu)
Publications
Funders
Fluid seepage at the continental margin offshore Co
Fluid seepage at the continental margin offshore Costa Rica
Division of Ocean Sciences (nsf.gov)
Division of Ocean Sciences (nsf.gov)
A hydrothermal seep on the Costa Rica margin: middle ground in a cor
Division of Ocean Sciences (nsf.gov)
Division of Ocean Sciences (nsf.gov)

Visualization of the connections of a DMP and related datasets, publications, funders, organizations and people

Implementation of Proof Of Concept

Having proven the value of a new metadata schema and ability to connect DMPs with research outputs via DMP IDs, the next phase was to have the DMPHub become an integral part of the DMPTool ecosystem. The DMPTool service was updated to implement features allowing researchers to provide more structured information about their intended research outputs and export DMPs as JSON defined by the common standard. A researcher can now, with a single button click, register their DMP in the DMPHub system, receive a DMP ID, and be presented with a public facing landing page for their data management plan displaying all of the networked components and outputs over time of the DMP.

Success Metrics

Assigning digital object identifiers (DOIs) to persistently identify DMPs is a trend that we have seen already. Since 2019, more than 200 DMPs have been assigned a DOI for their identification. In the one month DMPTool released it's feature to create DMP IDs, we have seen a 50% increase in this number.

Integrating with External Systems

With the infrastructure in place and initial feature development completed, in 2021 we will continue to release new features to expand the possibilities of the new networked DMP helping to ensure transparency in the research process and promoting good data practices for UC researchers. Many of these new features are currently being pilot tested as part of the FAIR Island Project where the University of California Gump South Pacific Research Station is a participating institution. Through the FAIR Island Project, by implementing mandatory registration requirements along with optimal data policies that make data findable, accessible, interoperable and reusable, machine-actionable DMPs will be utilized for tracking provenance, attribution, compliance, deposit, and publication of all research data collected on the island.

The team will continue to extend their work with the UC NRS and the Santa Cruz Island Reserve and integrations with the NRS's reservations management tool, which tracks data from projects and teams spanning scientific domains and institutions.

The DMPHub is a simple Ruby on Rails based API and is open source so anyone can deploy, customize and contribute to its ongoing development. API access can be granted to anyone to their own DMP IDs.