EAGER DMPRoadmap: Making Data Management Plans Actionable

Overview

The California Digital Library (CDL) will work with an extensive coalition of national and international collaborators to convert static data management plans (DMPs) into machine-actionable documents useful for structuring the course of research activities and communicating with other systems. Open data policies are proliferating worldwide and researchers are now required to submit DMPs with most grant proposals that describe the data they will produce and plans for sharing and preserving it. 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 form pose similar challenges for other stakeholders across the research ecosystem, e.g., funders who must monitor compliance manually. The proposed project will solve this problem by building out DMPRoadmap, a new, internationalized platform to reposition DMPs as true hubs of the networked research ecosystem.

The CDL will design, develop, and prototype machine-actionable DMPs based on community-generated use cases; field research on DMP and data management practices; and pilot projects conducted with disciplinary and institutional partners. The work plan encompasses iterative phases of developing, implementing, and testing with various stakeholder groups using an agile methodology. Specific use cases include implementing a set of common standards and exchange protocols for DMPs to enable information to flow between DMPs and existing research information systems (e.g., offices of research, data repositories, faculty profile systems); leveraging persistent identifiers (e.g., DOIs for articles and datasets, ORCID iDs for people) to trigger push/pull notifications across systems that enable stakeholders to plan resources, connect research outputs, automate reporting and monitoring, get credit, and promote data discoverability, reuse, and reproducibility; among others. An open framework ensures that solutions will be available to the global research community and not restricted to DMPRoadmap users.

Intellectual Merit

The intellectual merit of the proposed work is to provide a deeper understanding of how research activity develops from a grant award and how DMPs can be used to track this activity. The project will provide new tools to support researchers with managing their data effectively, as well as finding and reusing data generated by others to further their own work. By advancing best practices for data-intensive research across all disciplines the project will further the goals of the scientific endeavor.

Broader Impacts

The principal impact of the proposed project is to transform one component of the increasingly digital research enterprise, the DMP, into an actively updated and machine-actionable hub to document and disseminate the products of research activity. Converting free-text responses to funder requirements into dynamic, verifiable data feeds will vastly improve the entire toolchain for all stakeholders and increase the velocity and availability of information across all disciplines. The software outputs of the pilot will be shared publicly with an open source license so that the community can continue to enhance and reuse the technology, extending the scholarly cyberinfrastructure in a scalable manner. All supporting data collected during the project will be made fully available for human and machine consumption. Since DMPs are rapidly becoming a global phenomenon, the outputs of the project will be of great interest to the entire research community. The breadth of the impact extends into the future of DMP policies worldwide, as machine-actionability makes change easier and continuous improvements possible. The project will provide insight into how to maximize investment in institutions and infrastructures in a manner that achieves policy goals by providing public access to government-funded research, and advances the overall public good.

EAGER DMPRoadmap: Making Data Management Plans Actionable

1. Introduction

Data management plans (DMPs), written documents that describe practices and procedures around the data a researcher plans to acquire over the course of a research project, are the primary method by which funding agencies set expectations and measure compliance with policies related to the management and sharing of data. However, in their current form DMPs are static documents that cannot be updated or acted upon to achieve policy goals. The activities described in this proposal will transform DMPs into machine- as well as human-readable and actionable documents. The Data Documentation Initiative (DDI) defines machine-actionable as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure" [1]. The proliferation of international working groups focused on this vision of enhanced DMPs (e.g., Research Data Alliance [RDA] Active DMPs, Force11 FAIR DMPs) signals widespread interest in repositioning DMPs as living documents useful for structuring the course of research activities and integrating with other systems and workflows. This transformation will provide significant benefits to the complex ecosystem of stakeholders (institutions, funders, repositories, and others) that currently employ DMPs, but it will also require a significant degree of coordination to achieve practical results. Specifically, active DMPs will allow:

1. Institutions (e.g., academic libraries and their parent research universities) to provide effective, evidence-based data services;
2. Funders (e.g., National Science Foundation [NSF]) to monitor the data-related activities associated with individual grants;
3. Infrastructure providers (e.g., data repositories) to plan their resources; and
4. Researchers to manage, share, and discover data more easily.

Expectations related to how researchers should manage and share their data have a long history [2]. More recently, the explosive growth of data-intensive research coupled with concerns about the integrity of the research process [3–5] have prompted stakeholder groups including public and private funding agencies to set policies surrounding data. A 2013 memorandum from the White House Office of Science and Technology Policy instructed agencies with R&D budgets exceeding $100 million to ensure public availability of the products of federally funded research [6], and resulted in a majority of US agencies issuing requirements that a DMP be submitted as part of a grant proposal. Similar policies were already in place in the UK and now apply across Europe since the European Commission extended its Open Research Data pilot to cover all projects funded through Horizon2020, including a requirement for a DMP [7]. National data policies are in the planning stages for Canada, Argentina, South Africa, and Japan [8–9], among others.

Multiple tools have been developed to meet the skyrocketing demand for creating DMPs that comply with funder mandates. The cornerstone of support for DMPs in the US is the DMPTool [10], developed in 2011 by the California Digital Library (CDL) and founding collaborators—Digital Curation Centre (DCC), DataONE, Smithsonian Institution, University of California San Diego, University of California Los Angeles, University of Virginia, and University of Illinois at Urbana-Champaign—to provide an open, central clearinghouse of information about funder requirements, data management standards, and customized guidance. The CDL has continued to develop the DMPTool and track evolving funder requirements, maintaining 33 templates that address specific requirements for 17 different funders. The tool now serves over 23,000 users from 210 affiliated institutions and resulting in 18,700 plans. The DCC has parallel experience with DMPonline [11], which addresses UK and EU requirements and is now being adopted across Europe, Canada, and Australia. At this stage, we have two popular platforms that will soon become one openly developed tool: DMPRoadmap [12]. DMPRoadmap is a core piece of DMP

infrastructure that encompasses our respective services—both free to users anywhere in the world—in addition to national and organizational deployments [13]. This engaged and increasingly international user community (30+ total countries represented) has begun contributing directly to the DMPRoadmap codebase and provides insight into data policies and DMP requirements in diverse national and disciplinary contexts. Also based in the US, the Interdisciplinary Earth Data Alliance (IEDA) DMP Tool [14] allows researchers to generate a DMP for inclusion in NSF proposals, especially those related to solid earth data.

The NSF has been on the leading edge of DMP-related efforts and many other funders have modeled their requirements on those developed at the NSF. It is not surprising, therefore, that the majority of the scholarship examining perceptions and behaviors related to DMPs is based on those submitted with NSF proposals. Studies reveal that researchers exhibit a striking amount of variability in their understanding and application of important data management-related concepts like documentation (e.g., metadata), data reuse, and long-term preservation [15–17]. Furthermore, there appears to be no significant difference between how the DMPs of funded versus unfunded applications describe practices related to data storage and sharing [18–19]. Taken together, this research suggests that, while the current suite of DMP-related tools has enabled researchers to comply with DMP mandates, they have had a limited effect on achieving the goals of those mandates.

Active, machine-actionable DMPs present an opportunity to move DMPs beyond a compliance exercise by providing needed structure, interoperability, and added-value functionality to support open, reusable research data. In order to achieve this vision, the research community must first arrive at a shared understanding of the challenges and work together to create solutions in an open, community-driven manner [20]. We propose to leverage our experience, international contacts, and resources to design and develop an open framework for active DMPs.

2. Project Scope and Goals

The research project will explore the following questions:

- What information could be fed automatically from other systems into DMPs and vice versa?
- What actions could or should a DMP trigger?
- When should DMPs be updated?
- Who should know a DMP was updated?
- What role can DMPs play in education and training for data management?
- How can we offer tailored guidance that is relevant for a particular discipline?
- How can we bring DMPs closer to data and active research practices?
- What are the potential pathways for exposing DMPs and related outputs (e.g., publishing, depositing in a repository)?
- Which versions of a DMP should be archived and for how long?

We will investigate the feasibility of active DMPs through research and prototyping that builds on existing initiatives and infrastructure. This will enable us to be highly experimental and yet practical as our work will inform wider community efforts and draw on the experience of our broad coalition of national and international collaborators. To carry out this research we will:

1. Conduct field research via workshops at disciplinary conferences and campus-based events to explore the alignment of current research practices with active DMP workflows, and potentially reimagine the entire process of writing a DMP;
2. Enhance the DMPRoadmap platform to support active DMP use cases; and
3. Conduct DMP pilot projects with domain-specific and institutional stakeholders to test the free flow of relevant information across services and systems.

2.1. Field Research. Data management strategies can vary dramatically between and even within disciplines. Only a handful of communities have cohered around common practices and present opportunities for in-depth investigation of existing norms regarding DMPs and data management strategies. Therefore, we will focus field research and pilot projects on two of these broadly defined research communities: 1) environmental, oceanographic, and ecological researchers, and 2) biomedical researchers. We will conduct primary field research via workshops at disciplinary conferences (e.g., Earth Science Information Partners, American Geophysical Union, Intelligent Systems for Molecular Biology) and universities. This will enable us to address the question: what would make DMPs useful to researchers? In addition, we will document current practices, tools, and workflows employed by researchers to explore opportunities for integrating DMPs into the active research process. We will perform user testing of the enhanced DMPRoadmap platform to assess technical usability as well as researcher attitudes and perceptions toward active DMPs and what features and/or functionality they value. Workshop events will also allow us to evaluate the potential for more radical changes to the creation and use of DMPs (e.g., generating DMPs from information captured in an electronic lab notebook), that will ultimately require additional researcher and funder engagement.

2.2. Active DMP Use Cases. We have collected and prioritized a baseline of active DMP use cases. One of the major ways we did so was through hosting a workshop at the 12th International Digital Curation Conference (IDCC17) in Edinburgh, UK (Feb 20, 2017). The workshop participants included 47 people from 16 countries representing funders, researchers, developers, repository managers, university administrators, librarians, and other service providers. We collected additional inputs for the active DMP use cases through an informal survey of existing work (using social media channels, personal connections, and published literature) and direct participation in general and domain-specific working groups (e.g., RDA Active DMPs, Force11 FAIR DMPs, DDI metadata group for social science repositories, DataONE). We synthesized the outputs from these activities in a white paper [21–22], with a central theme of interoperability and exchange of information across research systems.

High-priority use cases include:

- Common data model: All stakeholders have expressed a need for common standards and protocols as a foundation to enable information flow between DMPs and systems in a standardized manner. We believe this can be achieved using existing standards and vocabularies. At the RDA 9th Plenary Meeting in Barcelona, Spain (Apr 6, 2017) we launched a working group to undertake this critical activity and define common standards for DMPs on a 12-month timeline. The group will define a common data model with a core set of elements that can be extended to follow best practices developed in various research communities to ensure broad adoption and enable interoperability of information contained in DMPs. We plan to implement the recommended standards in the DMPRoadmap platform, and will encourage similar efforts by other infrastructure providers who participate in the RDA. This presents a potential path for incorporating work from an ongoing project to customize the IEDA DMP Tool with a controlled vocabulary for NSF DMPs [23]. Once the standards are deployed, we will continue to receive and respond to community feedback.
- Building APIs (Application Programming Interface: a set of commands, functions, protocols, and objects that programmers can use to interact with external systems): APIs are one potential mechanism for achieving some level of funder integration to facilitate grant submission, monitoring, and reporting (i.e., by exchanging information between funders, offices of research, repositories, etc. programmatically). Automatic feeds from APIs could also help institutions and other service providers stay up to date with funder requirements for DMPs in order to maintain templates and offer appropriate support for researchers. In turn, this could help funders

demonstrate that DMP quality and compliance have an impact on funding success, which would contribute to improving the quality of DMPs and data management practices.

- Persistent identifiers (PIDs): A PID is a globally unique alphanumeric string assigned to a digital resource that provides a persistent link to that resource wherever it may live on the web. There are numerous PID schemes within the research data and scholarly publishing world: e.g., the Digital Object Identifier (DOI) is used for articles and datasets, ORCID iDs identify people, the Crossref Funder Registry is used for funder disambiguation, Research Resource IDs (RRIDs) are used for research objects (e.g. model organisms, software tools) [24–27], among others. Employing PIDs in DMPs would allow stakeholders to link people, grants, organizations, licenses, and associated research objects and track assertions made about them. These assertions would also enable DMPs to remain useful throughout the research process, converting them into a dynamic inventory of research activities that can trigger actions at the appropriate moment (e.g., automated status updates when a grant is awarded or dataset deposited; facilitating new collaborations and downstream uses of data and methods that are clearly documented in a single location).
- Repository recommender service: The majority of DMP requirements ask researchers to identify an intended data repository, but repositories rarely play an active role in the planning process. We have an extensive list of use cases for a repository recommender service to assist researchers with making an appropriate selection. This service could employ DataCite's re3data (a global registry of data repositories) [28] as well as community-curated lists such as BioSharing (a list of data standards, databases, and policies for the life, environmental, and biomedical sciences; based at the University of Oxford e-Research Centre) [29]. When a researcher names a repository in a DMP, the active DMP could alert repositories to data in the pipeline and enable them to initiate discussions with researchers early on. It would also facilitate capacity planning, allow for monitoring of changing user requirements, and increase the efficiency of the deposit process. A DOI or other identifier could be sent back to automatically update the DMP. This confirmation that data have been deposited as outlined in the DMP would be useful for funders, contributors, and institutions.

All of these examples help to ensure scientific integrity and reliable access to information. Moreover, actionable resource links allow the research objects, people, etc. to be tracked and cited enabling researchers to get credit for employing good data management practices and providing further incentives.

2.3. Pilot Projects. As part of an iterative process for developing, implementing, testing, and refining the use cases outlined above, we will model domain-specific and institutional pilot projects to assess what information can realistically move between stakeholders, systems, and research workflows, again focusing on the two target research communities. One pilot project involves partnering with the Biological and Chemical Oceanographic Data Management Office (BCO-DMO) [30] that services researchers funded through a variety of NSF programs to tailor guidance, structure plans, and automate processes in the DMPRoadmap platform. We will use DMPs from the GEOTRACES [31] project, a long-term, international study of marine biogeochemistry. For the biomedical pilot project, we will collaborate with BioSharing and the Wellcome Trust, a private biomedical research funder based in the UK; both are embedded in larger initiatives such as ELIXIR [32], a European life science infrastructure project with 21 member states and 180 participating research organizations.

2.4. Summary of Goals

Throughout the project we will report out and continue to engage with the research community through participation in working groups, conferences, email lists, and communication channels for our DMP services. We will test and refine the active DMP use cases via DMPRoadmap, and facilitate discussions

about how to prioritize our next steps as a community through use-driven experimentation in multiple directions. All software development for DMPRoadmap will continue to be conducted in an open, transparent manner, with the code and documentation available on GitHub under an open-source MIT license. DMPs have become a fundamental component of the increasingly digital research enterprise, and the global research community needs a basic framework with flexibility to experiment and develop their own solutions. Our pilot project will inform long-term infrastructure and training needs for DMPs and research data management writ large.

3. Strategic Partnership

The CDL possesses the requisite experience and expertise to provide solutions for active DMPs. At this point, the DMPTool has brought the CDL recognition among funders and a deep understanding of how researchers and funders use DMPs as part of their workflows. The new DMPRoadmap platform that will underpin the forthcoming versions (end of summer 2017) of the DMPTool, the DCC's DMPonline, and other national and organizational deployments addresses US-specific DMP requirements in a truly global research ecosystem. DMPRoadmap offers a central hub for testing by a highly distributed and engaged network of end users (including researchers, institutional administrators, and funders representing > 50,000 users in total); this will allow us to collect immediate feedback and make continual refinements to the design and deployment of active DMP functionality throughout this project. As a freely available, open-source community project, DMPRoadmap represents the best technology to move forward with the active DMP agenda to support scalability and maximize the impact of infrastructure investments.

In addition to offering technical solutions, we have built a vast social network, contribute actively to international working groups, and participate in disciplinary initiatives (e.g., DataONE, ELIXIR). CDL is part of the University of California system and has a deep connection with the research communities included in the pilot through a variety of additional curation tools, services, and research projects.

4. Project Outline

Our two-year project consists of five overlapping units:

- **Unit One** (Months 1-12): we continue conducting field research via workshops, with an emphasis on identifying researcher and funder attitudes toward DMPs and exploring ideas for making them a more useful exercise.
- **Unit Two** (Months 1-12): we enhance the DMPRoadmap platform by implementing community-developed use cases (e.g., common standards, APIs, PIDs, repository recommender).
- **Unit Three** (Months 6-18): we partner with stakeholders to conduct active DMP pilot projects that test the general use cases as well as develop additional domain- and context-specific use cases.
- **Unit Four** (Months 12-24): we refine the general use cases and implement new ones based on requirements gathered during Unit Three in DMPRoadmap.
- **Unit Five** (Months 4, 10, 16, 22): we conduct usability and user acceptance testing of the enhanced DMPRoadmap platform at regular intervals, make adjustments, and report out to the larger community through our standard communication channels and at conferences.

4.1. Unit One: Active DMP Field Research

We will use a mixture of qualitative and quantitative methods to evaluate user needs. Qualitative assessment will incorporate in-depth interviews (web-based or in person) to observe current RDM practices and workflows, gauge current perceptions and use of DMPs, and identify features that would enhance the value of DMPs. Quantitative assessment will include brief surveys (web- or paper-based; 5–10 minutes) that ask researchers and funders to identify the most desired features of active DMPs. The sample population will include researchers from the target domains and related sub-domains across

professional levels ranging from graduate students to principal investigators. We will also recruit from a range of educational and funding organizations (academia, museums, non-profits, public and private funders). Recruiting tools encompass social media (Twitter, blogs), email lists, and personal interactions at conferences, meetings, and university campuses and will leverage our connections with the UC system, DataONE, BCO-DMO, NSF, NIH, Wellcome Trust, Moore Foundation, among others.

4.2. Unit Two: Use Case Implementation

We will enhance the DMPRoadmap platform to support active DMPs, beginning with implementation of the four highest-priority use cases described in Section 2 and detailed in the white paper [21]. These general use cases address multiple stakeholders, especially core data infrastructure providers, and all fields of research. Recent refactoring of the codebase provides a robust foundation for experimentation, scalability to accommodate higher volumes of users, and extensibility of data and functionality via APIs and integrations with other systems. An open framework means that solutions will be available to the global research community and not restricted to DMPRoadmap users, a necessary condition for successful adoption of active DMPs.

1. Common standards: We will implement the recommendations of the RDA working group at the conclusion of their activities (estimated as April 2017). They will deliver a framework of common standards and exchange protocols (e.g., json) for DMPs to enable information to flow between plans and systems in a standardized manner. Requirements include making use of existing vocabularies and ontologies; support for new data types, models, and descriptions; linking to entities described with PIDs (e.g., people, repositories, licenses) for validation and scalability; and versioning to support actively updated DMPs.
2. APIs and integrations: In collaboration with existing partners, we will build APIs to create and retrieve plans in order to reduce administrative burdens (e.g., offices of research could pre-populate a DMP with certain pieces of information); embed external resources into the DMPRoadmap platform to offer better guidance (e.g., RDA Metadata Standards Directory, BioSharing); and export plans to external systems (e.g., institutional repositories, Dataverse, Figshare, Zenodo).
3. PIDs: We will incorporate additional PIDs into DMPRoadmap to trigger push/pull notifications across systems that enable stakeholders to plan resources, connect research outputs, automate reporting and monitoring, get credit, and promote discoverability, reuse, and reproducibility. We already support ORCID iDs and will add the Crossref Funder Registry, Organization IDs, RRIDs, and IDs for scientific protocols [33]. Assigning DataCite DOIs to DMPs of record (i.e., identified by offices of research as DMPs submitted with a grant proposal) presents another potential channel for passing information between systems.
4. Repository recommender service integrated with re3data.org: We will integrate with re3data to provide guidance for selecting a data repository that alerts researchers to eligibility requirements, metadata standards, etc., at the beginning of a project. We will also explore mechanisms to push notifications to repositories named in a DMP to inform them about data in the pipeline and support automated data tracking.

4.3 Unit Three: Disciplinary and Institutional Pilot Projects

The general active DMP use cases described in Unit Two present a baseline for identifying what information can realistically flow across DMPs and existing systems. In order to test the feasibility of the use cases we will conduct two disciplinary and two institutional pilot projects. Each pilot will involve analyzing DMPs from completed research projects and identifying what components can be structured and passed between systems, then recreating these DMPs with the enhanced DMPRoadmap tool and testing the machine-actionability of the resulting DMPs. We will leverage a variety of APIs, PIDs, and

notifications between repositories and other key systems implemented as part of Unit Two. The pilots offer one mechanism for measuring success as we evaluate whether the new DMPRoadmap features work in practice. This will potentially reveal points where active DMPs are constrained by and/or must be adapted to current infrastructure, systems, and practices.

1. Disciplinary pilots: BCO-DMO employs the DMPTool as part of their current data management toolchain, which encompasses all phases of the research lifecycle. The GEOTRACES project is a long-term, international marine biogeochemistry study managed through BCO-DMO and presents a compelling opportunity to work directly with researchers, data and repository managers, and funders to test the potential of active DMPs created with DMPRoadmap. For the biomedical research community we will identify a project supported by the Wellcome Trust and work collaboratively with other partners—NIH, BioSharing—to offer relevant guidance and automate the flow of information contained in DMPs across systems.
2. Institutional pilots: Purdue University and the UCSD will serve as institutional pilots to test the flow of information contained in DMPs across offices of research, libraries, repositories, and faculty profile systems. Both are power users and members of the DMPTool Steering Committee, with well-established networks within their respective institutional environments.

4.4. Unit Four: Use Case Refinement

During the second year, we will focus on refining the general use cases based on the results of Units One and Three. In addition, we will take any new requirements gathered in the course of Unit Three and implement additional use cases in DMPRoadmap.

4.5. Unit Five: Testing, Integration, and Presentation

The work plan encompasses iterative phases of developing, implementing, and testing with various groups using an agile methodology. We will work with the CDL User Experience (UX) team to perform user testing at regular intervals and report out to the community. By integrating the enhanced DMPRoadmap platform, community-generated use cases, and multiple rounds of testing we will gain a deeper understanding of how research activity develops from a grant award and how we can leverage DMPs to track this activity. Success will be determined by: 1) adoption by machines (i.e., consumers of APIs) expressed as an automated flow of information across systems and stakeholders, thereby alleviating the burden of manual processes; 2) increased adoption of DMPRoadmap by researchers and institutions; and 3) increased engagement with funders.

The CDL considers DMPRoadmap a key strategic initiative within its portfolio and remains committed to ongoing support. The DMPRoadmap software will be shared publicly with an open source license to enable and encourage community participation in its deployment, maintenance, and enhancement. All supplementary data collected during the project will also be made available for human and machine consumption.

5. Personnel Roles and Qualifications

Günter Waibel, UC Associate Vice Provost and Executive Director, CDL, will serve as PI and oversee all aspects of the project from setting priorities to ensuring successful completion. Waibel is a founding member of the DMPTool project from his previous role at the Smithsonian Institution and is deeply familiar with data management policies, infrastructure, and challenges from a variety of institutional perspectives.

Stephanie Simms, Research Data Specialist and DMPTool Service Manager, University of California Curation Center (UC3), CDL, will manage community interactions, requirements gathering, and development for DMPRoadmap. Simms will also work collaboratively with the CDL UX team to conduct

user testing and disseminate results. She is a co-chair of the RDA Active DMPs Interest Group and a member of the DataONE Users Group Steering Committee.

Brian Riley, DMPTool Technical Lead, UC3, CDL will lead all technical aspects of the project. He is currently the technical lead for the DMPTool and a core member of the DMPRoadmap team, which includes managing external contributions from the larger developer community and collaborating with project partners to design and build APIs.

6. Suitability for EAGER

The project we are proposing is highly experimental with a high payoff potential. The proposal is likely too risky for existing NSF programs as the community needs preliminary evidence to accept the validity of active DMPs. We aim to leverage DMPRoadmap as the proving ground with an initial set of community-generated and tested use cases. Furthermore, the proposal is focused on building discipline-agnostic infrastructure and lies outside the domain of general NSF grants. The resulting enhancements to the open-source DMPRoadmap application will be ready for production use and available for linking up with other systems in the research information ecosystem (e.g., researcher, funder, institution, data repository, etc.). The pilot implementation of active DMPs created with DMPRoadmap will demonstrate how these linkages can support research activity as it develops from a grant award. It also stands to produce other novel insights that facilitate training and outreach efforts to improve research data management practices and advance the research enterprise.

7. Broader Impacts

The transition to truly networked research relies on the ability of our scholarly community to work together to create outputs that are both human readable and machine actionable. This requires reimagining and retooling each component of our workflows and a long-term commitment on the part of stakeholders. To have the most impact, we must look to key touch points where we have the attention of all players (researchers, funders, academic institutions, data repositories, publishers), such as when a researcher plans a research project, proposes a grant to a funder, submits a research article to a publisher, or deposits data in a repository. The principal impact of our project will be to transform one key component, the DMP, into an active and machine-actionable hub to document and disseminate the products of research activity. Converting free-text responses to funder requirements into dynamic, verifiable data feeds will vastly improve the entire toolchain for all stakeholders and increase the velocity and availability of information across all disciplines.

The software outputs of the pilot will be shared publicly with an open source license so that the community can continue to enhance and reuse the technology, extending the scholarly cyberinfrastructure in a scalable manner. All supporting data collected during the project will be made fully available for human and machine-consumption. Since DMPs are rapidly becoming a global phenomenon, the outputs of the project will be of great interest to the entire research community. The breadth of the impact extends into the future of DMP policies worldwide, as machine-actionability makes change easier and continuous improvements possible. The project will provide insight into how to maximize investment in institutions and infrastructures in a manner that achieves policy goals by providing public access to government-funded research, and advances the overall public good.

8. Summary

The very notion of what a DMP could and should be must continue to evolve to provide appropriate support for the integrity, efficiency, and efficacy of the research enterprise. As the research community makes great strides forward in adopting and innovating new research data management best practices, the technical infrastructure underlying those practices must keep pace. Today’s static DMP documents

are inadequate to the challenge of underpinning a vibrant networked data management ecosystem coordinating the activities of all stakeholders. Our proposed enhancement of DMPRoadmap to support machine-actionable DMPs as active hubs for integrated research activities is a tangible first step towards the networked future, and will provide a firm basis for subsequent exploration of highly integrated, added-value research data services.

9. No prior NSF support has been received in the past five years.

References

[1] Data Documentation Initiative Alliance website. www.ddialliance.org/taxonomy/term/198. Accessed 27 May 2017.

[2] Weinberg, AM et al. 1963. Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information. A Report of the President's Science Advisory Committee. Washington, DC: U.S. Government Printing Office.

[3] Hey T, S Tansley, and KM Tolle. 2009. The fourth paradigm: Data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft Research.

[4] Ioannidis, JPA. 2005. Why Most Published Research Findings Are False. PLOS Medicine 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

[5] Ioannidis, JPA. 2014. How to Make More Published Research True. PLOS Medicine 11(10): e1001747. https://doi.org/10.1371/journal.pmed.1001747

[6] Holdren, JP. 2013. Memorandum for the heads of executive departments and agencies: Expanding public access to the results of federally funded research. Executive Office of the President, Office of Science and Technology. Available: https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

[7] European Commission. 2016. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 July 2016: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[8] Barsky, E, L Laliberté, A Leahey, and L Trimble. 2017. Collaborative Research Data Curation Services: A View from Canada, in Curating Research Data: Practical Strategies for your Digital Repository, Vol. 1, LR Johnston, ed: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

[9] Argentine Law 26.899: Creation of Digital Institutional Open Access Repositories, Personal or Shared. Ministerio de Ciencia, Tecnología e Innovación Productiva. http://repositorios.mincyt.gob.ar/recursos.php

[10] DMPTool website. https://dmptool.org/. Accessed 27 May 2017.

[11] DMPonline website. https://dmponline.dcc.ac.uk/. Accessed 27 May 2017.

[12] DMPRoadmap GitHub repository. https://github.com/dmproadmap. Accessed 27 May 2017.

[13] DMPRoadmap Local Installations Inventory. https://github.com/DMPRoadmap/roadmap/wiki/Local-installations-inventory. Accessed 27 May 2017.

[14] IEDA Data Management Plan Tool website. http://www.iedadata.org/compliance/plan. Accessed 27 May 2017.

[15] Bishoff, C, and L Johnston. 2015. Approaches to data sharing: An analysis of NSF data management plans from a large research university. Journal of Librarianship and Scholarly Communication 3(2): eP1231. http://doi.org/10.7710/2162-3309.1231

[16] Dietrich D, T Adamus, A Miner, and G Steinhart G. 2012. De-mystifying the data management requirements of research funders. Issues in Science and Technology Librarianship 70(1). http://dx.doi.org/10.5062/F44M92G2

[17] Parham, SW, J Carlson, P Hswe, B Westra, and A Whitmire. 2016. Using Data Management Plans to Explore Variability in Research Data Management Practices Across Domains. International Journal of Digital Curation 11(1): 53-67. http://dx.doi.org/10.2218/ijdc.v11i1.423

[18] Mischo WH, MC Schlembach, and MN O'Donnell. 2014. An analysis of data management plans in University of Illinois National Science Foundation grant proposals. Journal of eScience Librarianship 3(1): 31-43. http://dx.doi.org/10.7191/jeslib.2014.1060

[19] Van Tuyl, S, and AL Whitmire. 2016. Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia. PLOS ONE 11(2): e0147942. https://doi.org/10.1371/journal.pone.0147942

[20] Bourne, PE, JR Lorsch, and ED Green. 2015. Perspective: Sustaining the big-data ecosystem. Nature 527: S16–S17. https://doi.org/10.1038/527S16a

[21] Simms, S, S Jones, D Mietchen, and T Miksa. 2017. Machine-actionable data management plans (maDMPs). Research Ideas and Outcomes 3: e13086. https://doi.org/10.3897/rio.3.e13086

[22] Simms S, and S Jones S. 2017. Machine-actionable data management plans (maDMPs): IDCC17 workshop materials. Zenodo https://doi.org/10.5281/zenodo.475054

[23] National Science Foundation Award Abstract No 1649703. EAGER: Collaborative Research: Supporting Public Access to Supplemental Scholarly Products Generated from Grant Funded Research. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1649703&HistoricalAwards=false. Accessed 27 May 2017.

[24] Paskin, N. 1999. The digital object identifier system: Digital technology meets content management. Interlending and Document Supply 27(1): 13-16.

[25] Haak, LL, M Fenner, L Paglione, E. Pentz, and H Ratner. 2012. ORCID: A system to uniquely identify researchers. Learned Publishing 25(4): 259–264. https://doi.org/10.1087/20120404

[26] Meddings, K. 2013. FundRef: Connecting research funding to published outcomes. Insights 26(3): 272–276. http://doi.org/10.1629/2048-7754.98

[27] Bandrowski, A, M Brush, JS Grethe, MA Haendel, DN Kennedy, S Hill, PR Hof, ME Martone, M Pols, SC Tan, N Washington, E Zudilova-Seinstra, and N Vasilevsky. 2016. The Resource Identification Initiative: A cultural shift in publishing. Journal of Comparative Neurology 524(1): 8-22. https://dx.doi.org/10.1002/cne.23913

[28] re3data.org Registry of Research Data Repositories website. http://www.re3data.org/. Accessed 27 May 2017.

[29] BioSharing.org website. https://biosharing.org/. Accessed 27 May 2017.

[30] Biological and Chemical Oceanographic Data Management Office. http://www.bco-dmo.org/. Accessed 27 May 2017.

[31] GEOTRACES: An International Study of the Marine Biogeochemical Cycles of Trace Elements and Their Isotopes website. http://geotraces.org/. Accessed 27 May 2017.

[32] ELIXIR Europe website. https://www.elixir-europe.org/. Accessed 27 May 2017.

[33] Teytelman L, A Stoliartchouk, L Kindler, and BL Hurwitz. 2016. Protocols.io: Virtual Communities for Protocol Development and Discussion. PLoS Biology 14(8): e1002538. https://doi.org/10.1371/journal.pbio.1002538

Data Management Plan
EAGER DMPRoadmap: Making Data Management Plans Actionable

Types of data produced

Three types of data will be generated during this project:

1) Qualitative and quantitative survey results collected as part of Unit One will be captured in paper forms and via electronic methods (e.g., Qualtrics). Paper surveys will be digitized by hand and and originals will be kept for the duration of the project. All survey data will be compiled as csv files. We will comply with Institutional Review Board policies and secure approval before conducting the surveys with human subjects. Simms and Jones will be responsible for all survey data.
2) Software produced or modified as part of the project will be maintained in the DMPRoadmap GitHub repository during and after the project (https://github.com/DMPRoadmap/roadmap), where it is available under an open source MIT License. Riley will be responsible for managing the repository and related documentation.
3) Miscellaneous research products such as additional use cases developed as part of Unit Three, community feedback, project status communications, etc. will be disseminated publicly throughout the project via the project website, blogs, peer-reviewed articles, and data repositories as appropriate. Communication channels and resulting products will be managed by Simms and Jones.

Data and metadata standards

The DMPRoadmap software produced by the project will conform to accepted community best practices, including version control (using Git) with tagging of major releases, a permissive open-source license (MIT License), public availability in a community repository (GitHub), inline comments, reference and tutorial documentation with download and installation instructions that is available from within the software and from a community website, and extensive test coverage.

Policies for access and sharing

All three types of data produced in the course of this project will be publicly available for review, evaluation, and use as they are generated during the project and after its completion. Announcements about software and data availability will be made using a variety of channels (e.g., blogs, Twitter, email lists) targeting all interested stakeholder communities. All survey data will be anonymized to remove personally identifiable information in compliance with IRB protocols for human subjects research.

Policies for re-use, redistribution

All software products resulting from this project will be reusable and redistributable during the project and after its completion. The only restriction placed on redistribution of the software is that the copyright and license statement be kept intact as required by the MIT open source license. The software is expected to be of interest to national and international data infrastructure providers, data centers and repositories, institutional administrators, and individual researchers.

Plans for archiving and preservation

DMPRoadmap is an active software product and will continue to be managed in the community GitHub repository. We will document current and future releases with release notes and the code for each version will be publicly available using Git history. Major versions of the software will be deposited in the Merritt Repository Service and findable with a persistent identifier.

All survey data will be deposited to Dash, a data publication platform, and preserved in the Merritt Repository Service to provide public access and long-term storage upon completion of the project. All

data will be publicly available upon deposit and findable through a DataCite DOI granted by Dash. The datasets will be maintained for as long as they are of continuing value to researchers and project collaborators, for a minimum of five years.

Dash and Merritt are both managed by the University of California Curation Center (UC3). Merritt relies on a highly fault tolerant microservices architecture with significant redundancy of all computational and storage components and has not experienced any data loss over its seven years of production operation.