# Roadmap: A Research Data Management Advisory Platform

## Executive summary

The DMPTool and DMPonline were developed to meet an emerging need arising from the advent of open data policies and each is now well established as *the* resource for researchers seeking guidance in creating data management plans (DMPs) in the US and UK respectively. Both services, and their sponsoring organizations, the California Digital Library (CDL) and the Digital Curation Centre (DCC), have succeeded in enabling researchers to comply with funder requirements in producing DMPs. However, this is just one step along the road to advancing open science.

We see an opportunity to further leverage DMPs to support open science by integrating them into the broader ecosystem of data management infrastructure. In order to achieve this goal, we must redefine success to include not just adoption of our services by institutions but also widespread adoption by individual researchers, disciplinary communities, and funders. Working together with all stakeholders to make DMPs an essential, open part of the research lifecycle, and not just a matter of compliance, is the next step toward effectively managing and sharing research data.

We propose to join forces and build a new, global data management advisory platform that links DMPs to other components of the research lifecycle. The biomedical research community provides an opportunity to adapt the infrastructure and associated educational resources to one specific disciplinary community and plug into new initiatives. We will reposition DMPs as living documents useful for structuring the course of biomedical research activities and integrating with related data management systems to lower the barriers for implementation and promote culture change. Consolidating around a single platform for DMPs extends our reach, keeps costs down, and moves best practices forward, allowing us to participate in a truly global open science ecosystem.

## Problem Statement

Research data management (RDM) that enables open science is now acknowledged as a global challenge: research is global, policies are becoming global, and thus the need is global. Successful strategies for meeting this need require coordination of all efforts—infrastructure and education—at the global scale. In order to be discoverable, accessible, and reusable in ways that advance science, data need to be managed properly from the outset. Data management plans (DMPs) are already part of the policy landscape and we see an opportunity to leverage them to support open science by integrating them into the broader ecosystem of data management infrastructure. DMPs are also a useful tool to educate researchers and promote beneficial culture change.

While data science principles are beginning to appear in higher education curricula, many researchers remain unaware of the evolving norms of scientific best practices, and the specific tools and guidance that may be available to them through institutional and disciplinary affiliations. Worthwhile RDM efforts are underway by individual early adopters and champions, but the true promise of open science will occur only once these activities are broadly accepted and practiced by the entire community.

Researchers typically identify and affiliate more strongly with their discipline rather than their host institution; this underscores the importance of community-focused efforts to develop RDM standards and best practices. As providers of data management planning services, we have succeeded in supporting institution-based efforts but now seek closer engagement with individual disciplinary communities. The Open Science Prize offers an opportunity to work with the biomedical research community as a pilot for integrating DMPs with established RDM infrastructure, researcher workflows, and other community initiatives (e.g., bioCADDIE[1], BioSharing[2], and ELIXIR[3]).

The positive impact from successful RDM adoption would be far reaching. According to the most recent published information, the US National Institutes of Health (NIH) alone made over 52,000 grant awards to 35,000 principal investigators totaling $24.3 billion in the past FY 2015.[4] In the UK, the Medical Research Council (MRC) awarded 336 grants totaling £243.2 million (FY 2014),[5] while the Wellcome Trust awarded 1,227 totaling £497.7 million (FY 2013).[6] Similar patterns of public funding are found in many other national jurisdictions. Since many of these projects undoubtedly involve large research teams, this represents an enormous wellspring of transnational research activity with potential value for advancing global public health. The various outputs of that research—publications as well as data, software, and workflows—deserve to be affirmatively managed, preserved, published, and (re)used in order to justify public spending, catalyze scientific understanding and advancement, incentivize innovation and exploration, and ensure the integrity of the scientific enterprise. Education and access to relevant tools are crucial components of more widespread RDM adoption and more open scientific practices in biomedical research and beyond.

The CDL and DCC have achieved important successes in first-generation service offerings providing public access to RDM guidance and resources. The CDL's DMPTool[7] and the DCC's DMPonline[8] focus specifically on supporting US and UK researchers in fulfilling their data management planning obligations arising from funder mandates, pre-publication requirements, and institutional policies. The intention of the Roadmap project is to converge on a common technical platform for proffering RDM advice and supporting open science. The new platform will combine all existing functionality from the two tools regarding the DMP use case. More importantly, it will reposition DMPs as living documents useful for structuring the course of research activities and integrating with related data management systems and workflows. We see greater potential for the DMP as a dynamic checklist for pre- and post-award reporting; a manifest of research products that can be linked with published outputs; and a record of

[1] https://biocaddie.org/
[2] https://biosharing.org/
[3] https://www.elixir-europe.org/
[4] National Institutes of Health (2016), *NIH Awards by Location & Organization*. https://report.nih.gov/award/
[5] Medical Research Council (2015), *Annual Report and Accounts, 2014/2015*. http://www.mrc.ac.uk/publications/browse/annual-report-and-accounts-2014-15/
[6] Wellcome Trust (2016), *Wellcome Trust Grants Awarded 2013/14*. http://www.wellcome.ac.uk/Managing-a-grant/Grants-awarded/
[7] http://dmptool.org/
[8] https://dmponline.dcc.ac.uk/

data, from primary through processing stages, that could be passed to repositories. The DMP will therefore not only support the management of the data but boost its discoverability and reuse.

The CDL and DCC have collaborated informally from the start of our independent DMP activities to reduce needless duplication of effort and share experiences. For the Roadmap project, we will establish a formal partnership to pool our resources and leverage past investment towards a common goal of actualizing the greater potential for DMPs. Our organizations are well positioned to accomplish this goal with our deep knowledge of the technical as well as the community aspects of RDM in different national contexts. The Roadmap system and services based on it will be applicable to any national jurisdiction or international research community. Our work already supports the US and UK domains and is being adopted in many other countries; a single system offers a single point of interoperation and an opportunity to extend our reach.

## The Roadmap Project

New work on our respective systems is already underway to enable internationalization, integrate with other organizations and technical platforms, and encourage greater openness with DMPs. By joining forces, the Roadmap system will consolidate these efforts and move beyond a narrow focus on specific funders in specific countries, and even beyond institutional boundaries, to create a framework for engaging with disciplinary communities directly. These critical stakeholder groups have access to the appropriate social networks and domain expertise to set standards and coordinate effective training and outreach activities. The biomedical research community as well as the funders that support their research have already made significant investments in RDM infrastructure and policies; this community presents an exciting opportunity to repurpose DMPs as a mechanism for connecting existing systems and evolving initiatives throughout the full research lifecycle. Here we outline ideas for extending existing DMP infrastructure to support biomedical researchers with sound data creation and management from start to finish, thereby maximizing the potential for data availability and reuse.

### *Internationalization*

Both the CDL and DCC are engaged in international initiatives. By formalizing a partnership to co-develop one system, we will signal to the global research community that there is one place to create DMPs and find advisory information. This action extends our reach and impact by consolidating DMP efforts within and across communities at an international scale. In the same manner, we will continue opening up access to the service and DMPs created with it in order to advance an international agenda for effective RDM and open science. At present, our combined reach for institutional users extends throughout the US, UK, Canada, much of Europe, and into South Africa, Australia, and Singapore, with new inquiries coming in daily. Unaffiliated users include researchers throughout the developing world.

DCC has already secured a grant from the University of Edinburgh Innovation Fund to develop locale-aware support for DMPonline[9]. The need for this functionality stems from increasing diversity in the DMPonline user community. Planned enhancements for localization include tailoring the language used to other features that align the system for user communities within a specific locale (e.g., date and time conventions, data format conventions, funder lists, DMP templates, user authentication procedures). Locales can range from national contexts to funders, institutions, or research consortia that span multiple boundaries. In the context of biomedical research, we would offer a menu of appropriate funder templates and plug into community standards issued through organizations such as BioSharing and ELIXIR. We also plan to refine our guidance and training materials in consultation with international user groups, beginning at the 2016 International Digital Curation Conference[10].

## *Data Reuse*

The Roadmap system will incorporate feature enhancements that enable DMPs to be implemented and ultimately promote data sharing and reuse. For example, the new system will support a repository recommendation service for NIH-hosted[11] and other relevant repositories, including those listed by the Wellcome Trust[12]. This will help integrate DMPs into workflows for all stakeholders. It is necessary for researchers to understand repository requirements from the outset in order to plan effectively and increase the likelihood of successful data deposit and reusability. By the same token, data repositories would benefit from being able to predict future data deposits.

We will also implement a common metadata schema for DMPs to enable interoperability with other systems. The Consortia Advancing Standards in Research Administration Information (CASRAI)[13] group has adapted the DCC Checklist for a DMP[14] and the set of themes used for guidance in DMPonline as a starting point for their vocabulary. The next step is to create mappings from templates to concepts defined in the CASRAI list[15]. The Open Science Prize will allow us to identify and incorporate ontologies specific to biomedical, health, and other life science data emanating from various working and interest groups (as well as the Research Data Alliance Active DMP Interest Group).

The CDL is currently enhancing the DMPTool API to integrate with the Open Science Framework[16] developed by the Center for Open Science[17]. This integration pilot will enable researchers to access supporting documentation about research data (i.e., DMPs) across platforms, which represents another

[9] https://dmponline.dcc.ac.uk/files/DMPonline-v4-LocaleSupport.pdf
[10] http://www.dcc.ac.uk/events/idcc16
[11] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
[12] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/WTX060360.htm
[13] http://casrai.org/
[14] http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf
[15] http://dictionary.casrai.org/DMP_Ethics_Review
[16] https://osf.io/
[17] https://cos.io/

step toward breaking free from silos, opening up data and systems. Additional integration projects that promote data reuse might include SHARE, the Crossref/Datacite DOI Event Tracking system, EUDAT, Zenodo, Jupyter notebooks, and researcher profile systems such as VIVO. We will identify other possibilities in consultation with the biomedical research community.

Persistent identifiers are another important element for integrating systems, workflows, and research outputs to produce a record of the data a project will make public. Both tools already support ORCIDs, but we plan to add support for other identifiers such as FundRef[18] for funding agency identification. Another CDL service for creating and managing persistent identifiers, EZID[19], is currently working with Identifiers.org[20] and the NIH to convert GenBank accession numbers into persistent identifiers. In the context of biomedical research we could also include identifiers for things such as instruments and reagents. Linking systems and research outputs across the web increases the chances that data will be discovered, accessed, and (re)used.

## *Assessment*

Enabling interoperability will transform the DMP from a static text file to a dynamic tool for planning and assessment. We aim to convert the DMP into an index of where data and other outputs are being collected to assure that open data policies will be enforceable by alleviating the burden of manual compliance checks and reporting for funders and researchers alike. Increasing compliance in turn increases transparency and openness in research practices.

## *Openness*

We are already encouraging researchers to publish open DMPs with public sharing features in the DMPTool. We will carry this over into the merged Roadmap system, in addition to exploring other avenues for elevating the status of DMPs as valuable research products. One possibility is to change the default plan visibility to "public" instead of "private." Another value-added service would be to assign DOIs to DMPs to encourage further sharing—of data and plans, as the latter represent a record of the data a project will make public—and enhance discoverability. An export option to the Zenodo repository is planned as part of the DCC activities in OpenAIRE to automatically assign DOIs to published DMPs. We have also been in contact with the Research Ideas and Outcomes (RIO)[21] journal about publishing exemplary DMPs.

Another measure of openness is expanding our ability to reach areas that do not have the resources to support data management on their own. Both of our systems are being used by unaffiliated researchers in developing countries for proactive planning within their national contexts as well as to comply with

[18] http://www.crossref.org/fundingdata/registry.html
[19] http://ezid.cdlib.org/
[20] http://identifiers.org/
[21] http://riojournal.com/

DMP and open data requirements issued by international funding bodies. Open science has a global agenda, and by making DMPs true infrastructure in a global open access community we will elevate research and open data for reuse.

## Phase I Development Targets

At present, both the DMPTool and DMPonline are open source projects available on GitHub, with free hosting and support provided by their sponsoring organizations. The Roadmap system will also be available in a new GitHub repository under an MIT license.

We are already outlining our co-development process and partnership agreement, beginning with a gap analysis of the two systems and roadmap consolidation. In 2016 Q2 we will begin adding features and anticipate our first coordinated release with a single product team by 2017 Q1. New features will include:

- extending authentication and localization support to all instances
- identifying partners and issuing an integration roadmap for external/reporting systems
- formalizing the concept of themes in DMPonline into an actionable data model for pan-funder requirements

Throughout this process, we will maintain outreach and training programs in our national contexts and continue coordinating outreach efforts to the international community. As part of these efforts, we will organize meetings with funders and researchers to evaluate current practices and workflows and determine additional points of integration with existing systems, metadata requirements, etc. In addition to the sponsors of the Open Science Prize, we plan to consult with ELIXIR, OpenAIRE, EUDAT, BioSharing, and bioCADDIE as we design these meetings.

## Phase II Development Targets

If we are successful with the Phase I prize, we will move forward with a second coordinated product release that incorporates additional features and integrations, and pursue additional funder/researcher events.

## Enabling Open Science

There is tremendous potential in removing silos to create intuitive workflows and connect services for data management activities. We propose to build a new, global framework for data management planning that links DMPs to researchers, funders, publications, data, and other components of the research lifecycle. By refocusing our efforts from promoting the creation of static DMPs to comply with funder requirements to supporting the creation of high-quality, dynamic DMPs that can be implemented and used as a structural hub for subsequent research activities, we will further enable the open science revolution. Consolidating around a single DMP platform extends our reach, keeps costs down, and moves best practices forward, allowing us to participate in a truly global open science ecosystem.