IJDC | *General Article*

Next-Generation Data Management Plans: Global Machine-Actionable FAIR

Stephanie Simms
California Digital Library

Sarah Jones
Digital Curation Centre

Abstract

At IDCC 2016 the Digital Curation Centre (DCC) and University of California Curation Center (UC3) at the California Digital Library (CDL) announced plans to merge our respective data management planning tools, DMPonline and DMPTool, into a single platform. By formalizing our partnership and co-developing a core infrastructure for data management plans (DMPs), we aim to meet the skyrocketing demand for our services in our national, and increasingly international, contexts. The larger goal is to engage with what is now a global DMP agenda and help make DMPs a more useful exercise for all stakeholders in the research enterprise. This year we offer a progress report that encompasses our co-development roadmap and future enhancements focused on implementing use cases for machine-actionable DMPs.

Received 28 November 2016 ~ *Accepted* 21 February 2017

Correspondence should be addressed to Stephanie Simms, University of California, Office of the President 415 20th Street, 4th Floor, Oakland, CA 94612-2901 Email: Stephanie.Simms@ucop.edu

An earlier version of this paper was presented at the 12th International Digital Curation Conference

http://dx.doi.org/10.2218/ijdc.v12i1.513
DOI: 10.2218/ijdc.v12i1.513

Introduction

The open scholarship revolution has progressed to a point where top-level policies mandate open access to the results of government-funded research, including research data, in the US, UK, and EU, with similar principles and policies gaining momentum in Australia, Canada, South Africa, and elsewhere. DMPs are the primary vehicle for complying with these policies, and because research is a global enterprise, awareness of DMPs has spread throughout the research community. Journal and institutional data policies that overlap with funder policies are also proliferating, which only increases the need for standards, training, and infrastructure that enable DMPs to be implemented in order to produce FAIR data (Findable, Accessible, Interoperable, Reusable)[1] at the end of a project. As the wheels of culture change turn slowly and steadily toward greater openness, the time has arrived to leverage DMPs to achieve these policy goals. The next generation of DMPs must convert them from an administrative exercise and instead produce dynamic inventories of digital research methods and outputs that evolve over time. We still need a human-readable narrative, but underneath the DMP could have more thematic, machine-actionable[2] richness with added value for all stakeholders. Here we present a brief summary of recent developments in open data that impact DMPs and related services, followed by an update on our project to build an internationalised DMP infrastructure (codenamed DMP Roadmap[3]) that will support the creation of machine-actionable DMPs.

Truly global efforts to support open scholarship are underway, in part owing to changes in both policy and practice over the past year. The European Commission extended its Open Research Data pilot to cover all work areas as of January 2017. Projects funded under Horizon 2020 will require a DMP in which they are encouraged to explain how the data will be FAIR, and are expected to be "as open as possible, as closed as necessary" (European Commission, 2016). Initiatives in the UK have also sought to harmonise the existing policy landscape, with a multi-stakeholder group representing various interests in the higher education sector agreeing a Concordat on Open Research Data in 2016 (HEFCE et al., 2016). Also in 2016, the US established the Interagency Working Group on Open Science (IWGOS) to coordinate the response to the 2013 Office of Science and Technology Policy (OSTP) memorandum; the memo directed all federal agencies with over $100 million in annual research and development expenditures to provide public access to publications as well as data from funded projects, including a requirement for DMPs. The IWGOS closed out the year with a progress report indicating that all 22 agencies required to do so have released their public access plans, now available on the newly launched CENDI.gov website[4]. Canada continues to prepare for future implementation of the 2015 Tri-Agency Principles, a draft statement on federally funded data management policy that includes DMPs and open data. The Canadian Association of Research Libraries (CARL) Portage network[5]

1 FAIR Principles: https://www.force11.org/fairprinciples
2 The Data Documentation Initiative defines machine-actionable as "information that is structured in a consistent way so that machines, or computers, can be programmed against the structure." See: www.ddialliance.org/taxonomy/term/198
3 DMP Roadmap: https://github.com/dmproadmap
4 See: https://cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html
5 CARL Portage Network: https://portagenetwork.ca/

has taken a proactive approach with several pilot projects designed to coordinate RDM infrastructure, expertise, and services at the national level (Barsky et al., 2017).

On top of the accumulation of national data policies, 2016 ushered in a series of related developments in openness that contribute to the machine-actionable DMP conversation. To name a few:

- More publishers articulated clear data policies, e.g. Springer Nature Research Data Policies[6] apply to over 600 journals.
- PLOS now requires an ORCID for all corresponding authors at the time of manuscript submission to promote discoverability and credit.
- The Gates Foundation reinforced support for open access and open data by preventing funded researchers from publishing in journals that do not comply with its policy[7], which came into force at the beginning of 2017; this includes non-compliant high-impact journals such as *Science*, *Nature*, *PNAS*, and *NEJM*.
- Researchers throughout the world continued to circumvent subscription access to scholarly literature by using Sci-Hub (Bohannon, 2016).
- Library consortia in Germany and Taiwan cancelled (or threatened to cancel) subscriptions to Elsevier journals because of open-access related conflicts, and Peru cancelled over a lack of government funding for expensive paid access (Schiermeier and Rodríguez Mega, 2017).
- Reproducibility continued to gain prominence, e.g. the US National Institutes of Health (NIH) Policy on Rigor and Reproducibility[8] came into force for most NIH and AHRQ grant proposals received in 2016.
- The Software Citation Principles (Smith et al., 2016) recognized software as an important product of modern research that must be managed alongside data and other outputs.

This flurry of open scholarship activity, both top-down and bottom-up, across all stakeholders continues to drive adoption of our services. DMPonline and the DMPTool were developed in 2011 to support open data policies in the UK and US, respectively, but today our organisations engage with users throughout the world. An upsurge in international users, especially from Europe, Australia, South Africa, Latin America, China, and the Middle East, is evident from email addresses for new accounts and web analytics. In addition, local installations of our open source tools, as both national and institutional services, continue to multiply[9].

Over the past year, the DMP community has validated our decision to consolidate our efforts by merging our technical platforms and coordinating outreach activities. The DMPRoadmap project feeds into a larger goal of harnessing the work of international DMP projects to benefit the entire community. We are also engaged with some vibrant international working groups (e.g., Research Data Alliance Active DMPs, FORCE11 FAIR DMPs, Data Documentation Initiative DMP Metadata group) that have provided the opportunity to shift our focus to developing use cases for machine-actionable DMPs.

6 Springer Nature Research Data Policies: http://www.springernature.com/gp/group/data-policy/
7 Gates Foundation Open Access Policy: http://www.gatesfoundation.org/How-We-Work/General-Information/Open-Access-Policy
8 NIH Policy on Rigor and Reproducibility: https://grants.nih.gov/reproducibility/index.htm
9 See the DMPRoadmap GitHub wiki for a complete list: https://github.com/DMPRoadmap/roadmap/wiki/Local-installations-inventory

Update on Co-Development

In 2016 we consolidated our project team and our plans for the merged platform, and we are now testing a co-development process that will provide a framework for community contributions down the line. The new platform is based on the DMPonline codebase and incorporates recent internationalisation work by the Portage network in Canada. Following a gap analysis, we issued a roadmap[10] that incorporates existing functionality from the DMPTool and enhancements in a few key areas. Priorities include:

- API development
- More robust institutional branding
- OAuth link for ORCID
- Shibboleth support through eduGAIN
- New and improved text editor: Substance Forms[11]
- Public DMPs list[12]
- Lifecycle to indicate the status of plans
- Flag for test plans to exclude them from usage statistics
- Redesign of plan creation wizard
- Copy plan option for creating new plans
- Copy template option for admins
- Template export
- Plan review and commenting features
- Enhanced admin controls

Data Model and Restructuring

One major addition to the co-development roadmap was a revision of the data model and detailed restructuring exercise. In Autumn 2016, heightened usage of the DCC-hosted DMPTuuli[13] service in Finland revealed serious limitations to performance of the DMPonline code. The Finnish consortium had achieved adoption by the Academy of Finland and the tool was recommended for use in the main Autumn call for funding proposals. In the weeks preceding the deadline, 130–160 new users were signing up each day. On three occasions, the volume of simultaneous use caused the tool to grind to a halt. The system was working throughout but page load times were so slow that it became unusable. The DCC devised a temporary fix to support users during the call period. This included some minor code changes to reduce the number of database queries and an increase in memory and CPUs. It was clear that major restructuring was needed though to ensure future performance and scalability; this work took priority in late 2016–early 2017.

10 DMP Roadmap: https://github.com/DMPRoadmap/roadmap/wiki/Development-roadmap
11 Substance Forms: http://substance.io/
12 cf. the DMPTool Public DMPs: https://dmptool.org/public_dmps
13 DMPTuuli: https://www.dmptuuli.fi/

The performance issues related to both the database structure and code: the code was not accessing the database in an efficient manner, and the data model was complex and difficult to understand. We ran through a series of use cases to simplify the data model and redesigned the database schema to only include natural entities. Next we wrote and tested migrations to transfer existing data to the new schema. An additional step involved refactoring the code to reduce the number of queries and to isolate data access from the views. Prior to these changes the system was getting blocked as multiple processes were waiting on the database. Certain areas of the tool, such as the plan writing page with multiple windows for each question and a locking mechanism to avoid overwriting, were particular candidates for optimisation.

The data model is now much more intuitive, which will make it easier for external contributors to work on DMPRoadmap. The improved documentation and streamlined architecture also lay the foundation for integrations with other systems and delivery of DMP information via APIs, a key priority in our machine-actionable plans.

Harnessing Community Development Efforts

Tuesday mornings in California/afternoons in Scotland now involve weekly check-in calls for sprint planning and review. As we test and refine our processes, we continue to add documentation to the GitHub repository where the code is available under an MIT license. In addition, we publish monthly progress updates on the DCC[14] and DMPTool blogs[15] The UX team at the CDL is assisting with the implementation of new features and improving the usability of some existing functionality. They will contribute to larger redesign efforts following the release of the new platform. In addition, we are consulting with the UC Berkeley Web Accessibility team to ensure that the new platform is accessible for users with disabilities. We estimate that DMPonline will be running the Roadmap code by the time of IDCC 2017 and the DMPTool will migrate soon thereafter.

The international developer community – chiefly the Portage developer responsible for the DMP Assistant[16] service and the Inist-CNRS developer who runs DMP OPIDoR[17] – has already contributed back to the project and offers a test case for incorporating external work that adds value for all users. We are working together to develop guidelines for contributors (available in draft form on GitHub[18]). We will also begin holding a monthly call for external developers (in Canada, European countries, and elsewhere) to provide a forum for sharing immediate and future needs, and ideas for customisations and enhancements. Drawing together these distributed DMP projects will avoid duplication of effort and consolidate value upstream to ensure maximum benefits for everyone.

Vision for the Future: Machine-Actionable DMPs

Last year we outlined possible approaches to advancing the DMP agenda including mining DMPs to understand and report current practices; promoting interoperability through community standards for DMPs; and integrating our platform with other

14 DCC Blog: http://www.dcc.ac.uk/drupal/blog
15 DMPTool Blog: https://blog.dmptool.org/
16 DMP Assistant: https://assistant.portagenetwork.ca/
17 DMP OPIDoR: https://dmp.opidor.fr
18 See: https://github.com/DMPRoadmap/roadmap/wiki/Developer-guide

research data systems to facilitate institutional processes (Simms et al., in press). Various data management stakeholders have been discussing these ideas for some time now and we determined that a core DMP infrastructure would enable us to implement them, contributing to a shared vision of machine-actionable DMPs that enable:

- Institutions to manage their data
- Funders to mine the DMPs they receive
- Infrastructure providers to plan their resources
- Researchers to discover data

We have made significant progress in surveying existing work in this area and crafting use cases that will reposition DMPs as living documents useful for structuring the course of research activities and integrating with other systems and workflows. We will continue to present in international fora and participate in working groups (e.g. RDA, FORCE11, DDI, CASRAI) to develop the use cases further through community engagement. So far the use cases encompass a controlled vocabulary for DMPs; integrations with other systems (e.g. Zenodo, Dataverse, Figshare, OSF, PURE, grant management systems, electronic lab notebooks); passing information to/from repositories; leveraging persistent identifiers (PIDs); and building APIs. We have also replaced the current text editor with Substance Forms that comes with built-in structuring and commenting features. The new editor offers a better user experience in the short term and lays the foundation for exploring a completely different approach to authoring DMPs through future collaboration with the Substance team. Here we outline our evolving ideas for use cases and pilot projects to make DMPs machine-actionable.

DMPonline Themes

The first use case that will be deployed in the new system involves extending the DMPonline themes to all templates worldwide. Themes represent the most common topics addressed in DMPs and work like tags to associate questions and guidance. Questions within DMP templates can be tagged with one or more themes (e.g., Data Volume, Data Format), and guidance can be written by theme to allow organisations to apply their advice over multiple templates at once. This alleviates the need for monitoring changes in requirements and updating guidelines each time a new template is released. It also represents an opportunity to test the potential of this basic structure for identifying sections of text to mine, e.g. to identify a repository named in a DMP and the volume of data in the pipeline. We evaluated the existing set of 28 themes and guidance developed by the DCC to address UK funder requirements in the context of US and EU requirements, and then reduced the list to 14 themes to help streamline guidance and avoid confusion (for admins using the themes as well as end users consuming guidance while writing a plan). The new set of themes incorporates feedback from user communities on both sides of the pond and is available on GitHub[19].

Next steps involve using the themes to craft a more specific DMP vocabulary in collaboration with key stakeholders such as the Data Documentation Initiative (DDI) working group, the RDA and FORCE11 DMP groups. We will also track related initiatives for open government data, such as the World Bank metadata for research projects.

19 DMPRoadmap Themes: https://github.com/DMPRoadmap/roadmap/wiki/Themes

Repository Use Cases

Repository use cases are a high priority since the selection of a data repository requires up-front planning, especially with regard to storage and human labour costs, and affects every phase of a project's lifecycle. Within this broad category, possible use cases include the following:

1. A repository recommender service integrated with re3data.org that provides researchers with eligibility requirements, metadata standards, etc. at the beginning of a project.

2. Push notifications to repositories named in a DMP. This includes critical details such as the volume and types of data in the pipeline. It also facilitates communication between researchers and repository managers at the beginning of a project.

3. Support for automated data tracking using PIDs (see below), e.g. push/pull notifications when a dataset is deposited in a named repository. This allows repositories, institutions, and funders to monitor compliance and automatically populates CVs, researcher profile systems, reports, etc. with information about research outputs.

Persistent Identifiers (PIDS)

PIDs are a key ingredient for enabling information to pass between existing systems and workflows to plan resources, connect outputs, and automate reporting and monitoring. One possible implementation is to assign a Digital Object Identifier (DOI) to the DMP of record (i.e., the version of a DMP submitted with a grant proposal). This DOI provides a hook to dynamically update the DMP over time, e.g. with award details and other PIDs such as ORCIDs, FundRef IDs, Research Resource IDs[20] for key biological resources, DOIs for articles, datasets (Starr et al., 2015), etc. In the case of Horizon 2020, project-level data could be provided by OpenAIRE so researchers could automatically populate administrative details when starting a new plan. This would help to associate the DMP with other project outputs and assist in compliance monitoring. For other funders where the DMP is required at grant application stage (e.g., ZonMw in the Netherlands, Academy of Finland), feeding the grant ID into the tool will identify active projects and automatically trigger the DMP onto the next stage. By incorporating PIDs, we can transform the DMP into a living document that captures everything from the planning to preservation and reporting phases of a research project.

API Use Cases

Three existing APIs support the creation of plans (e.g. Offices of Research can pre-populate a plan with the correct template, guidance, and basic details to reduce the burden on researchers), retrieval of custom guidance, and the generation of basic usage statistics. We are designing additional APIs to deliver other key information within plans and exchange data between research systems. CRIS systems, for example, hold relevant information about researchers and projects that could be fed into DMPs, and extensions could interoperate with other commonly used platforms like Pure. It is also

20 Resource Identification Portal: https://scicrunch.org/resources

useful for institutions, funders, and others to be able to harvest information and statistics to inform capacity planning, service requests, and provide consultation support. We are in the process of identifying priorities and gathering requirements for API development.

Pilot Projects for Machine-Actionable DMPs

As part of an iterative process for developing, implementing, testing, and refining these use cases, we will model domain-specific and institutional pilot projects to determine what information can realistically move between stakeholders, systems, and research workflows. One of the pilot projects involves partnering with the NSF-funded Biological and Chemical Oceanography Data Management Office (BCO-DMO)[21] to use its GEOTRACES corpus[22], a long-term, international study of marine biogeochemistry. We are also collaborating with key stakeholders in the biomedical research community – the Wellcome Trust, NIH, BioSharing, ELIXIR – to identify a second pilot project supported by one of these groups. Purdue University has agreed to serve as an institutional pilot to model the flow of information across Offices of Research, libraries, repositories, and faculty profile systems. In addition to technical solutions, these projects will expand our capacity to connect with key stakeholders, with particular emphasis on better addressing the needs and practices of researchers and funders.

Promoting DMPs as Open Resources

In a related effort, we are designing workflows to promote the idea of DMPs as public, open, discoverable resources. So far this includes enhancing the public DMPs list already supported in the DMPTool and inviting authors of exemplary DMPs to publish them in a RIO Journal collection[23]. By working together to achieve machine-actionable and publicly available DMPs, we can move beyond static text files to create a dynamic inventory of digital research methods, protocols, environments, software, articles, data, and other related outputs with mutual benefits for all stakeholders. There now seems to be a critical mass and willingness to move toward a networked, machine-actionable approach to data management.

Conclusion

Our work already supports the US and UK domains and is being adopted in many other countries. So far ca. 30,000 users have created ca. 40,000 plans with our tools since they were launched in 2011. Our combined insight into DMPs as a fundamental part of modern research practice and our focus on researchers as the primary stakeholders make this project uniquely positioned to move DMPs forward. A single system offers a single point of interoperation and an opportunity to promote open scholarship at a global scale.

21 BCO-DMO: http://www.bco-dmo.org/
22 GOTRACES: http://www.geotraces.org/
23 RIO Journal Collection: http://riojournal.com/browse_user_collection_documents.php?collection_id=3&journal_id=17

Acknowledgements

Thanks to Beth Plale, Inna Kouper, and the Alfred P. Sloan Foundation for an RDA Data Share Fellowship awarded to S. Simms to develop use cases for machine-actionable DMPs. We are grateful to our global DMP user communities, especially Weiwei Shi and Chuck Humphrey at Portage and Benjamin Faure and Marie-Christine Jacquemot at Inist-CNRS who have been willing guinea pig contributors and excellent collaborators. Anna Gazdowicz and Lucy Greco from the UC Berkeley Web Access Team graciously donated their time and expertise. Tomasz Miksa (SBA Research, University of Vienna), Fernando Aguilar (IFCA-CSIC), the DDI DMP metadata working group, among others, have contributed inspiration and substance to the machine-actionable DMP use cases, which are absolutely a community-drive effort.

References

Barsky, E., Laliberté, L., Leahey, A., & Trimble, L. (2017). Collaborative research data curation services: A view from Canada. *Curating Research Data: Practical Strategies for your Digital Repository, 1.* Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf

Bohannon, J. (2016). Who's downloading pirated papers? Everyone. *Science, 352*(6285). doi:10.1126/science.aaf5664

European Commission. (2016). H2020 programme: Guidelines on FAIR data management in Horizon 2020, Version 3.0. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Fallaw, C., Dunham, E., Wickes, E., Strong, D., Stein, A., Zhang, Q., Rimkus, K.R., Ingram, W., & Imker, H.J. (2016). Overly honest data repository development. *The Code4Lib Journal* 34. Retrieved from http://hdl.handle.net/2142/91684

HEFCE, RCUK, Universities UK & Wellcome Trust. (2016). Concordat on open research data. Retrieved from http://www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf

Schiermeier, Q. & Rodrîguez Mega, E. (2017). Scientists in Germany, Peru, and Taiwan to lose access to Elsevier journals. *Nature, 541*(13). doi:10.1038/nature.2016.21223

Simms, S., Strong, M., Jones, S., & Ribeiro, M. (in press). The future of data management planning: Tools, policies, and players. *International Journal of Digital Curation.*

Smith A.M., Katz D.S., Niemeyer K.E., & FORCE11 Software Citation Working Group. (2016). Software citation principles. *PeerJ Computer Science* 2:e86. doi:10.7717/peerj-cs.86

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., et al. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science* 1. doi:/10.7717/peerj-cs.1