Open Access to Research Artifacts: Implementing the Next Generation Data Management Plan

ABSTRACT

The National Science Foundation began requiring a Data Management Plan – two pages of free text – to be included with research proposals in 2011. We describe a new vision for a Data Management Plan (DMP) that incorporates controlled vocabularies and semantic descriptions of the scholarly objects to be produced by the proposed project. We implement this vision in an open-source prototype web-based DMP tool, called ezDMP, at ezdmp.org. The integrated use of structured information in ezDMP permits several important goals. First, with minimal additional effort, researchers can create DMPs with more complete information on the scholarly objects to be produced. Second, research funders can productively query this structured information to learn about repository use and other patterns of scholarly objects creation. Finally, ezDMP puts a structure in place that can support the integration of information about digital scholars objects, in an organized and systematic way, into an emerging research data management environment.

KEYWORDS

Data Management Plan; Data Sharing; Code Sharing; Cyberinfrastructure for Research; Data Policy; Code Policy; Interoperability; Reproducible Research; Digital Repositories; Open Access; Scholarly Communication.

ASIS&T THESAURUS

Computer Systems: Interfaces; Information Utilities; Public Policy; Government Agencies; Digital Repositories; Authors; Standards Developing Organizations.

INTRODUCTION

Data Management Plans have been a required part of a National Science Foundation (NSF) proposal submission since 2011 and concern the planned/proposed artifact output of research grants. Artifacts can refer to datasets, software, workflow information, samples and other products of the research beyond the discoveries themselves. Reflecting on the seven years that Data Management Plans (DMPs) have been required, we describe a next generation Data Management Plan structure that serves the two principal DMP goals: first, to communicate and encourage awareness in the research community regarding priorities and modalities for artifact sharing, reuse, and research reproducibility; and second, to enable funders and community stakeholders to learn about research artifact creation, archiving, and reuse practices by researchers and other stakeholders.

The current NSF Data Management Plan guidelines limit the length of a free text document to two pages. Each of the seven directorates within the NSF provide domain specific guidance for the content of these two pages (e.g. which research artifacts should be discussed). This guidance raises awareness in the community but does not give specifics on the factors the DMP should address regarding artifact sharing. This can leave many crucial questions unaddressed such as details regarding artifact sharing such as licensing and terms of use, and artifact access, ownership, and stewardship, and repository use. We address this goal directly through the use of structured DMPs that prompt the researcher to (often optionally) address these issues. A DMP that is structured in this way permits machine readability and the extraction of information by the funding agency and submitting research institution. In this way, the next generation DMP permits funders to answer crucial questions such as: What are the patterns in repository use in research communities for the different types of artifacts? How do communities differ in archiving and sharing practices? Where are there gaps in existing infrastructure and support for research artifact sharing? Do completed research projects meet the goals stated in their DMPs? Under current funding agency DMP requirements answering these meta questions is next to impossible for the agencies, since DMPs are often submitted as freeform text documents.

In this article, we first outline and motivate a next generation DMP that enables funders to meet the two goals discussed, and then we present an implementation of a web-based interface that facilitates the straightforward production of such DMPs by researchers, librarians, and proposal writers.

OTHER DATA MANAGEMENT PLAN EFFORTS

Online tools that assist with the creation of the Data Management Plans that accompany research proposals are not a new idea. The DataONE project and the California Digital Library have created tools and many university libraries provide services in the creation of Data Management Plan for their researchers (Shreeves, 2014). There are DMP tool efforts in Europe, for example DMP Online (Sallans *et al.*, 2012) and the DMPTuuli project in Finland (Ahokas *et al.*, 2017). All these efforts, to our knowledge, do not use controlled vocabularies nor structured information in a template form, although in some cases the user can download and complete a docx template on their own. The IEDA DMP Tool (see https://www.iedadata.org/dmp/), is a structured webform geared primarily toward earth and ocean scientists. We build on and extend the IEDA efforts by implementing a structured process for gathering information and completing the DMP using controlled vocabularies, as described below.

THE NEXT GENERATION DATA MANAGEMENT PLAN

Over the last decade computing has become central to virtually all discoveries emanating from the scientific research enterprise. The vast majority of fields have embraced and leveraged data, computing power, and digital resources to advance and accelerate discovery (Donoho *et al.*, 2009). With these changes, the reliance on customized software for discovery has become commonplace across the research community, and the use of cyberinfrastructure tools and platforms for research has become standard with some research groups contributing tools themselves. An early example is the federally funded Wavelab software toolbox emerging from Stanford University that was widely adopted and set research and dissemination standards within the wavelets research community in signal processing (Buckheit *et al.*, 1995). This example indicates the importance and impact of documenting and sharing scholarly objects in a disciplined way, including details on the data, software, and research tools that were used to generate research findings, called "really reproducible research." (Claerbout & Karrenbach, 1992). Since then, reproducibility has become a topic of great research and policy interest today (see e.g. the Congressionally mandated National Academies of Sciences, Engineering, and Medicine's forthcoming consensus report "Reproducibility and Replicability in Science" at http://sites.nationalacademies.org/dbasse/bbcss/reproducibility_and_replicability_in_science/index.htm). Conversely, a lack of transparency regarding the computational implementation of the research hampers or even blocks efforts to reproduce or and verify results (Stodden, 2013). Recently steps toward enabling and rewarding the dissemination of the artifacts (e.g. data, code) that underlie published findings have been taken by journals (Stodden *et al.*, 2016) and institutions (AAU-APLU *et al.*, 2017). Similarly, a required Data Management Plan is a key part of an overall strategy by many funding agencies in facilitating the production of reproducible and transparent research findings (see e.g. https://www.nsf.gov/bfa/dias/policy/dmp.jsp and https://science.energy.gov/funding-opportunities/digital-data-management).

Many fields do not have established and widely adopted domain repositories, nor broadly agreed-upon metadata definitions for artifacts. This can create artifact interoperability issues and a lack of understanding of artifact provenance, including for example descriptions of data generation mechanisms. There is a wide range of possible artifact formats and a lack of community guidance on data release standards. Software also lacks generally accepted guidance on appropriate documentation and metadata standards. In addition, there is little guidance is on appropriate workflow information and information needed to, for example, use artifacts to regenerate published scientific results (Santana-Perez, 2017; Gil, 2011). This results in a situation where researchers may feel ill-equipped to meet DMP requirements.

Appropriate documentation for artifacts produced during the research along with a clear communication of how they underlie scientific results can enable reuse and accelerate discovery while reducing duplication of effort. The next generation DMP emerged via a community-driven NSF Advisory Committee Working Group.

Evolving the NSF Data Management Plan

The need to evolve the Data Management Plan was addressed by a Working Group of the NSF Advisory Committee on Cyberinfrastructure (ACCI) on "Data and Code Access and Reproducibility" formed in 2015, under Victoria Stodden's Committee co-chairship and with Helen Berman serving as Working Group chair. The Working Group produced a detailed set of recommendations for a DMP consistent with the NSF Public Access Plan that both communicated of the importance of research artifact dissemination to the community, and enabled analysis of DMPs by funders to improve understanding of artifact sharing patterns.

These recommendations were then implemented into a prototype web-based interface in 2018. To do this, we examined more than 1,350 anonymized data management plans in the IEDA DMP Tool to understand gaps, successes, and patterns of use. The reported research products from these DMPs fell into five categories: Software, Data Products, Curriculum, Physical Specimens, and Workflow Information. From our sample of DMPs, we compared and contrasted DMPs submitted to the different NSF Directorates. Finally, with the completion of a prototype ezDMP tool we surveyed potential users and presented the prototype to NSF program officers for feedback in 2018.

Communicating Artifact Dissemination Priorities

Prior to the completion of the prototype tool, the working group examined and collated information on all NSF DMP guidelines from the seven directorates. Although the high-level requirements are similar, the detailed requirements varied. After the ezDMP tool gathers basic demographic and proposal information such as the solicitation, a structured template is used for the five research product categories, as shown in Figure 1. The user can click through to an NSF Directorate's current published DMP guidance.

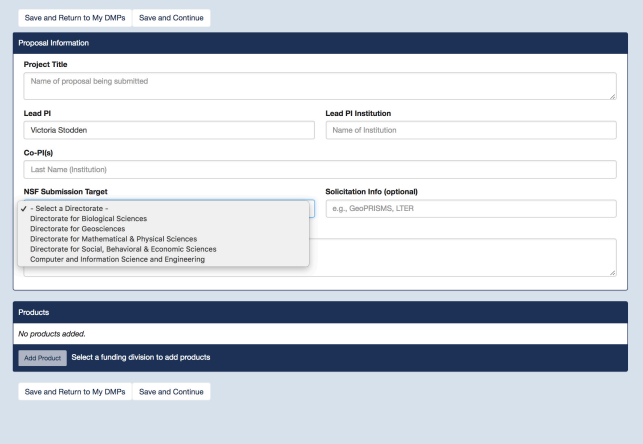

Figure 1. The ezDMP Data Management Plan is guided in the information it presents to the researcher guided by the DMP requirements specified by each NSF Directorate.

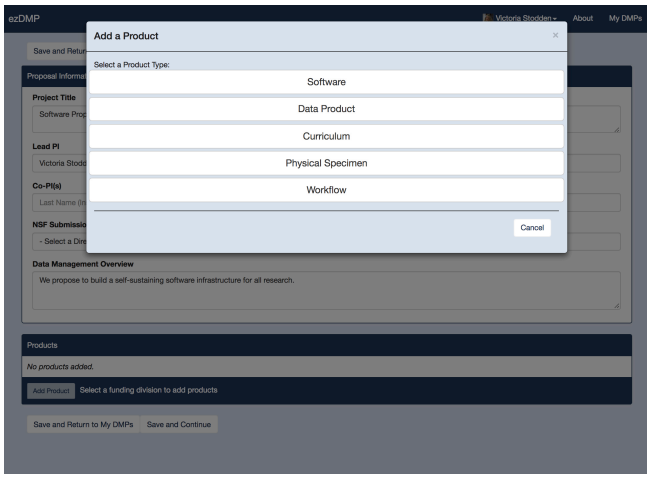

Figure 2. The addition of specific artifacts in ezDMP occurs in a structured way using controlled vocabularies.

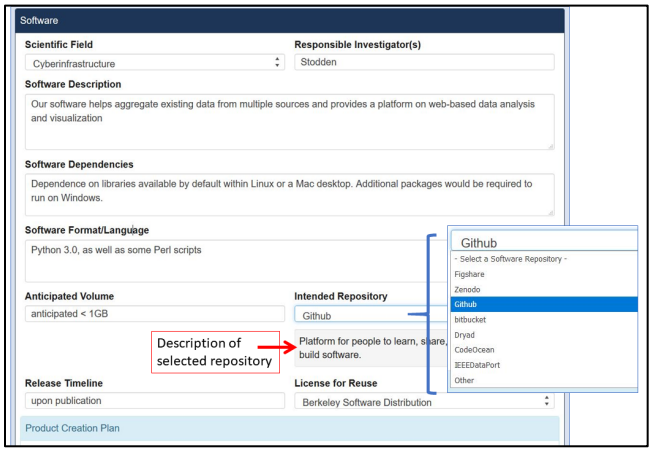

Figure 3. Repository choices for a software artifact. The interface also allows for information to be included in addition to that supplied in the drop-down menus, for example a repository not listed by the tool, so ezDMP can adapt to evolving community practices and funding agencies can learn about these changes in a systematic and timely way.

After completing demographic and solicitation information, the tool then presents the user with opportunities to enter information about each research artifact (dataset, software, curriculum materials, physical specimens, or workflow information) they expect to generate during the course of the project. For each artifact chosen, a structured set of choices are presented to elicit specialized information about the artifact with respect to attributes such as licensing, repository, stewardship, etc. As shown in Figure 3, at each stage the user always has the ability to enter information that does not currently appear in the choices presented by the template.

After completing the modules for the appropriate artifacts, a two-page pdf is returned to the author for inclusion in their funding proposal. It is possible for users to contribute descriptions of artifacts that may not currently exist in ezDMP and it is possible for free text to be added to any drop-down menu that describes artifacts. A new repository can be included this way, or other new modalities coming into use in the community, using text boxes for artifacts or descriptions that do not fit the template structure. In this way NSF can learn about artifacts and their requirements as they evolve over time. A novel contribution of the ezDMP tool is its communication to authors and researchers a list of potential repositories based on the type of artifact they will be producing. The tool also makes a second novel contribution by communicating information that should travel with artifacts, such as licensing and access information, which adds to the evolving discussion on Data Management Plans and reproducibility in the community.

Enabling the Study of DMPs (Learning from the Community)

The specific fields in the DMP template enable querying community practices in artifact sharing by funding agencies and institutional research offices. In the course of creating a DMP, information is collected on repository selection, licensing, NSF infrastructure and facility use, artifact formats and meta data, as well as information to use the artifacts and potentially reproducible the research results. The ezDMP tool also gathers information on planned artifact availability and retention. To do this, the ezDMP employs a controlled vocabulary that is specific to NSF Directorate and artifact type thereby enable data mining and an improved understanding of community practices.

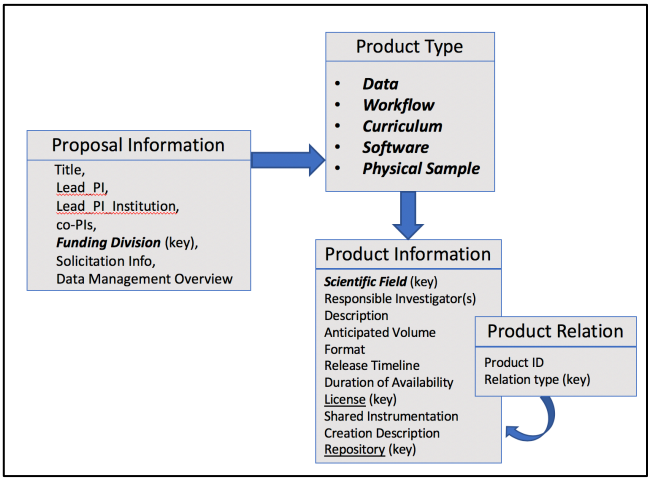

Figure 4. Conceptual schematic of the ezDMP Database Design showing the relationship between research artifacts and the use of controlled vocabularies when gathering information on artifacts produced by research grants. Fields in bold-italic control the options presented for underlined fields.

EZDMP: A WEB-BASED IMPLEMENTATION OF THE NEXT GENERATION DATA MANAGEMENT PLAN

As shown in Figure 4, information is gathered by the ezDMP web interface in a systematic way that preserves relationships between the information types. The ezDMP application was developed in Node.js using the Express.js framework with a PostgreSQL backend connected via object-relational mapping (ORM) and the pg-promise library. The front-end is built in Angular.js with fully responsive Bootstrap UI elements for desktop, mobile, and tablet support. User authentication is managed through

Google OAuth and ORCID, and user information is stored in JSON Web Tokens.

Back-end work included developing the database schema, populating and refining all necessary controlled vocabularies based on community input, and building all services necessary for desired functionality. The list of potential repositories is derived from curated repository lists we assembled. These repository lists are included in the back-end and enable the delivery of a menu of potential repositories to users based on division, product type and scientific field chosen. The ezDMP schema also accommodates relating artifacts to one another, such as data products that will be derived from software that will be developed.

While in the development phase, the application was made available for targeted testing by a variety of stakeholders. The web site with the prototype version of the ezDMP tool is https://www.ezdmp.org. The user interface source code is available at https://github.com/ezdmp/ezDMP-Site.

CONCLUSION

In this article, we have described the implementation of a next generation DMP and the motivation for the two key goals it addresses. These goals are to communicate policy priorities regarding artifact availability to the research community and to enable funders and community stakeholders to learn about research artifact creation, archiving, and reuse practices by researchers and other stakeholders. Our work has focused on the National Science Foundation and we note that other funding agencies are moving forward with Data Management Plans as well (see e.g. the October 2018 Request for Information by the National Institutes for Health entitled "Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research" https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html). We anticipate extending the tool to accommodate other funding sources in a customized way in the future. Within NSF, data and artifact policies are advancing, especially with respect to enabling reproducibility of results (see e.g. https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf and https://www.nsf.gov/cise/oac/ci2030/ACCI_CI2030Report_Approved_Pub.pdf).

We believe a next generation Data Management Plan, generated using a tool that produces a structured, machine readable, output using controlled vocabularies and semantic descriptions of the scholarly objects produced, will permit a greater understanding of practices regarding artifact creation, and availability, allowing for improved credit and recognition of these efforts. In addition, the approach of ezDMP will encourage greater development of artifact standards and interoperability by the research community and permit the incorporation of the Data Management Plan in a future data management environment. We see ezDMP is a first step toward realizing these goals.

ACKNOWLEDGMENTS

We thank the researchers and stakeholders who evaluated our tool and completed our feedback rubric. We also thank participants at RDA and other workshops for their valuable feedback. We are grateful for support from NSF awards 1649555, 1649545, and 1649703.

REFERENCES

AAU-APLU, Lynch, L., Nusser, S., Brown, S., Chasen, J., Dutta, D., . . . Wheeler, B. (2017). *AAU-APLU Public Access Working Group Report and Recommendations (White Paper)*. Retrieved from online: https://www.aau.edu/key-issues/aau-aplu-public-access-working-group-report-and-recommendations

Ahokas, M., Kuusniemi, M. E., & Friman, J. (2017). Tuuli project: accelerating data management planning in Finnish research organisations *International Journal of Digital Curation, 12*(2), 107-115.

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and Reproducible Research. In Antoniadis A. & O. G. (Eds.), *Wavelets and Statistics* (Vol. Lecture Notes in Statistics, pp. 55-81). New York, NY: Springer.

Claerbout, J. & Karrenback, M. (1992). Electronic documents give reproducible research a new meaning. In: Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics, pp. 601–604.

Donoho, D. L., Maleki, A., Shahram, M., Rahman, I. U., & Stodden, V. (2009). Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering, 11*(1), 8-18.

Gil, Y., Ratnakar, V., Kim, J., González-Calero. P. A., Groth, P., Moody, J., & Deelman, E. (2011). Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems*. 26(1).

Sallans, A., & Donnelly, M. (2012). DMP Online and DMPTool: Different Strategies Towards a Shared Goal. *International Journal of Digital Curation, 7*(2), 123-129. doi:10.2218/ijdc.v7i2.235

Shreeves, S. L. (2014). *Presenting the New and Improved DMPTool*. Paper presented at the Open Repositories 2014, Helsinki, Finland. http://hdl.handle.net/2142/49957

Santana-Perez, I., Ferreira da Silva, R., Rynge, M., Deelman, E., Perez-Hernandez, M. S., & Corcho, O. (2017). Reproducibility of Execution Environments in Computational Science Using Semantics and Clouds. *Future Generation Computer Systems*, 67, 354–367.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science, 354*(6317), 1240-1241.

Stodden, V., (2013). Resolving Irreproducibility in Empirical and Computational Research. *IMS Bull. Online*. http://bulletin.imstat.org/2013/11/resolving-irreproducibility-in-empirical-and-computational-research/