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

Victoria Stodden
School of Information Sciences
University of Illinois Urbana-Champaign
Champaign, IL USA
vcs@stodden.net

Vicki Ferrini
Earth Observatory of Columbia University
New York, NY USA
ferrini@ldeo.columbia.edu

Margaret Gabanyi
Research Collaboratory for Structural Biology
Rutgers University
Piscataway, NJ USA
gabanyi@rcsb.rutgers.edu

Kerstin Lehnert
Earth Observatory of Columbia University
New York, NY USA
lehnert@ldeo.columbia.edu

John Morton
Earth Observatory of Columbia University
New York, NY USA
jmorton@ldeo.columbia.edu

Helen Berman
Department of Chemistry and Chemical Biology
Rutgers University
Piscataway, NJ USA
berman@rcsb.rutgers.edu

ABSTRACT

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 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 about 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 research data management environments.

KEYWORDS

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

ASIS&T THESAURUS

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

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19 – 23 October, 2019

INTRODUCTION

Data Management Plans have been a required part of a National Science Foundation (NSF) proposal submission since 2011 and concern 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 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 collect data to enable funders and stakeholders to learn about research artifact creation, archiving, and reuse practices by researchers and others.

The current NSF Data Management Plan guidelines limit its length 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 unanswered such as artifact licensing and terms of use; artifact access, ownership, and stewardship; and repository use. We address this goal directly through the development of structured DMPs that prompt the researcher to (often optionally) address these pertinent issues. A DMP that is structured in this way permits machine readability and the automatic extraction of information. 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 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 requirements answering these questions is challenging for the agencies, since DMPs are submitted as freeform text documents.

In this article we outline and motivate a next generation DMP that enables funders to meet the two goals articulated above, and we present an implementation of a web-based tool that facilitates the production of such DMPs.

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 Plans 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). 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 template on their own. The IEDA DMP Tool (see https://www.iedadata.org/dmp/) is a structured webform geared 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 the scientific research enterprise. Fields have embraced and leveraged data, computing power, and digital resources to advance and accelerate discovery (Donoho et al., 2009). An early response to the increased need for transparency due to computation is described in (Buckheit et al., 1995), and includes details on sharing data, software, and research tools that were used to generate research findings, This practice was called “really reproducible research.” (Claerbout & Karrenbach, 1992; Stodden, 2013). Since then, reproducibility has become a topic of great research and policy interest (National Academies, 2019). Recently steps toward enabling and rewarding the dissemination of the artifacts that underlie published findings have been taken by journals (Stodden et al., 2016) and institutions (AAU-APLU et al., 2017). The Data Management Plan is a key part of an overall strategy by many funding agencies to facilitate the production of reproducible and transparent research (see https://www.nsf.gov/bfa/dias/policy/dmp.jsp and https://science.energy.gov/funding-opportunities/digital-data-management).

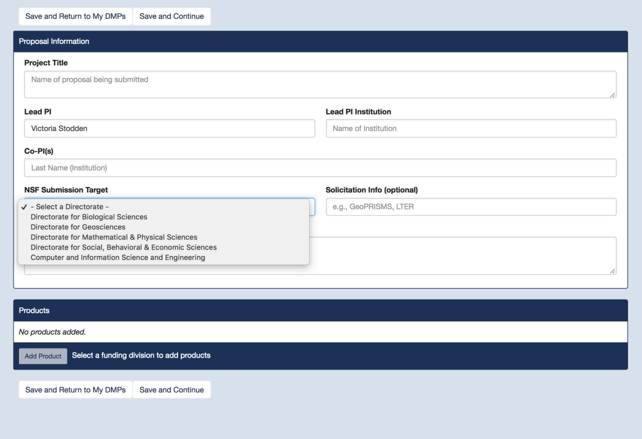
Many disciplines do not have established and widely adopted domain repositories, nor broadly agreed-upon meta data definitions for artifacts such as data and software. This can create artifact interoperability issues and gaps in artifact provenance and data generation mechanisms. There is a wide range of possible artifact formats and a lack of community guidance on artifact release 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). Therefore 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

A Working Group of the NSF Advisory Committee on Cyberinfrastructure (ACCI) called “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 (see https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf) 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, 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. The survey rubric is available at https://goo.gl/forms/CaEB3ddJ3iuUmpxS2.



Proposal Information

Project Title: Name of proposal being submitted

Lead PI: Victoria Stodden

Lead PI Institution: Name of Institution

Co-PI(s): Last Name (Institution)

NSF Submission Target: ✓ - Select a Directorate -

Solicitation Info (optional): e.g., GeoPRISMS, LTER

Products

No products added.

Select a funding division to add products

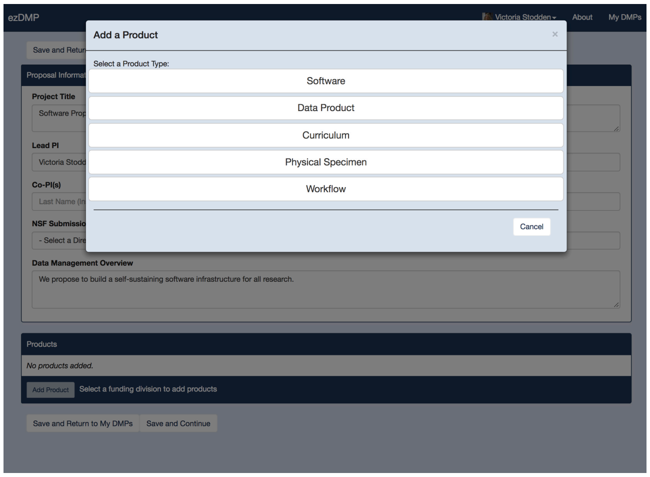
Save and Return to My DMPs
Save and Continue
- Select a Directorate -
Directorate for Biological Sciences
Directorate for Geosciences
Directorate for Mathematical & Physical Sciences
Directorate for Social, Behavioral & Economic Sciences
Computer and Information Science and Engineering
Add Product
Save and Return to My DMPs
Save and Continue

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.

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 highlevel requirements are similar, the detailed requirements varied. As shown in Figure 1, the ezDMP tool gathers basic demographic and proposal information from the user. The user can click through to an NSF Directorate's current DMP guidance.

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, as shown in Figure 2. A structured template is employed for the five research product categories.



Add a Product

Select a Product Type:

Project Title: Software Prop

Lead PI: Victoria Stodd

Co-PI(s): Last Name (In

NSF Submissio: - Select a Dire

Data Management Overview: We propose to build a self-sustaining software infrastructure for all research.

Products

No products added.

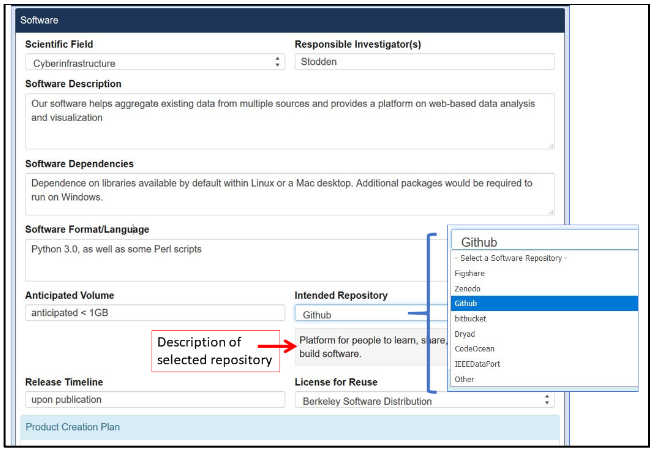
Add Product
Select a funding division to add products
Save and Return to My DMPs
Save and Continue

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

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 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. An example of a Data Management Plan produced using ezDMP is available at https://zenodo.org/record/3247756#.XQfZbdM3nOQ (Gabanyi, 2019).

It is possible for users to contribute descriptions of artifacts that may not currently exist in ezDMP and for free text to be added to any drop-down menu that describes artifacts. A new repository can be included this way, using text boxes for artifacts or descriptions that do not fit the current template structure. ezDMP can update its templates in this way and funders can learn about artifacts and their requirements as they evolve over time. A novel contribution of the ezDMP tool is its communication to users 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.


Software

Scientific Field: Cyberinfrastructure

Responsible Investigator(s): Stodden

Software Description: Our software helps aggregate existing data from multiple sources and provides a platform on web-based data analysis and visualization

Software Dependencies: Dependence on libraries available by default within Linux or a Mac desktop. Additional packages would be required to run on Windows.

Software Format/Language: Python 3.0, as well as some Perl scripts

Anticipated Volume: anticipated < 1GB

Intended Repository: Github

Description of selected repository

Platform for people to learn, share, build software.

Release Timeline: upon publication

License for Reuse: Berkeley Software Distribution

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.

Enabling the Study of DMPs (Learning from the Community)

The specific fields in the DMP template generate data that can be used to understand community practices in artifact sharing. In the course of creating a DMP, information is collected on repository selection, licensing, NSF infrastructure and facility use, artifact formats and meta data, and 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 specific to NSF Directorate and artifact type thereby enabling querying and information extraction.

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.

Back-end work included developing the database schema, populating and refining all necessary controlled vocabularies based on community input, and building 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. Data collected by the tool is archived internally.

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.

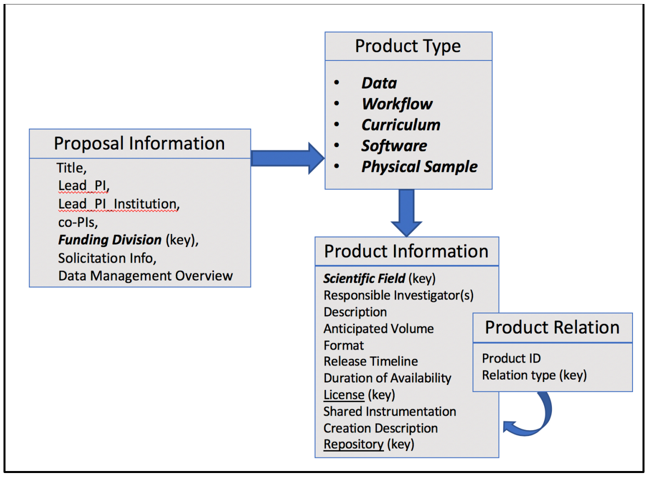


Figure 4. Conceptual schematic of the ezDMP Database Design showing the relationship between research artifacts and the use of controlled vocabularies when gathering information on artifacts produced by research grants. Fields in bold-italic control the options presented for underlined fields.

CONCLUSION

In this article, we have described the implementation of a next generation DMP and the motivation for the two key goals it addresses in facilitating greater access and transparency in research: 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.

Funding agencies are continuing to implement Data Management Plans (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). Recommendation 6-5 of the recent National Academies report on reproducibility exhorts the NSF to "[c]onsider extending NSF's current data-management plan to include other digital artifacts, such as software" (National Academies, 2019).

We believe a next generation Data Management Plan, generated using a tool that produces 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 pioneered by ezDMP will encourage greater development of artifact standards and interoperability and facilitate the incorporation of the Data Management Plan in future data management environments.

ACKNOWLEDGMENTS

We thank the stakeholders who provided feedback and completed our evaluation rubric. We also thank participants at RDA and other workshops for comments. NSF awards 1649555, 1649545, and 1649703 supported this work.

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: 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: *Proc. 62nd Ann. Intl Meeting Soc. 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.

Gabanyi, Margaret. (2019). Example ezDMP output (Version 0.1). Zenodo. doi: 10.5281/zenodo.3247756

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).

National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

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 *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/