Published on *Data Management Plan Tutorial* (https://www.e-education.psu.edu/dmpt)

Home > Modules

Introduction to Data Management Plans

What is a Data Management Plan (DMP)?

A data management plan is a document that tells how a researcher will collect, document, describe, share, and preserve the data that will be generated as part of a project.

Dr. Andrew Stephenson is Distinguished Professor of Biology and Associate Dean for Research and Graduate Education in the Eberly College of Science at Penn State. As an active researcher, he has generated and collected data for many years and served on many a panel reviewing grant proposals. From his perspective, data management plans make good sense. In the following video, he describes the elements of a DMP and why they are important.

Link to YouTube video. [1]

Many funding agencies are now requiring that grant applicants provide information about their data management plan (DMP) as part of their grant proposal. Since 2011 the National Science Foundation (NSF) has required researchers to include DMPs with their grant proposal applications.

Information about the data management plan requirement at the NSF website.
Credit: National Science Foundation [2]

DMPs are typically supplemental to a grant application. The NSF specifies that a plan should not exceed two pages. Other funding agencies may have different requirements for length; check with the guidelines of the grant program you are applying for. The NSF also understands that DMPs are not relevant for some projects. In such cases, the agency recommends that the researcher provides a statement explaining why a DMP is not being submitted.

NOTE: There are several directorates in the NSF that have more specific guidance than what follows in this tutorial. It is recommended that you refer to such guidelines (see the list in the Related Resources [3] tab above) if your directorate is included, in addition to taking this tutorial.

# Why Do You Need a Data Management Plan?

Obviously, the foremost reason for needing a plan is that agencies such as the NSF, the National Institute of Health (NIH), and the National Endowment for the Humanities (NEH) are requiring DMPs. Hear what Dr. Stephenson has to say about the impact of DMPs on choosing which grants to fund.

Link to YouTube video. [4]

There are other reasons, however, why formulating a plan for managing research data is important.

Reason One:

First, a DMP helps you plan and organize your data collection by having you think through the questions that will arise as you gather data. A DMP essentially documents key activities in the research data lifecycle, such as the collection, description, preservation, and access or discovery of data. Such documentation is crucial to reproducibility of research results which is a fundamental precept of scientific investigations.
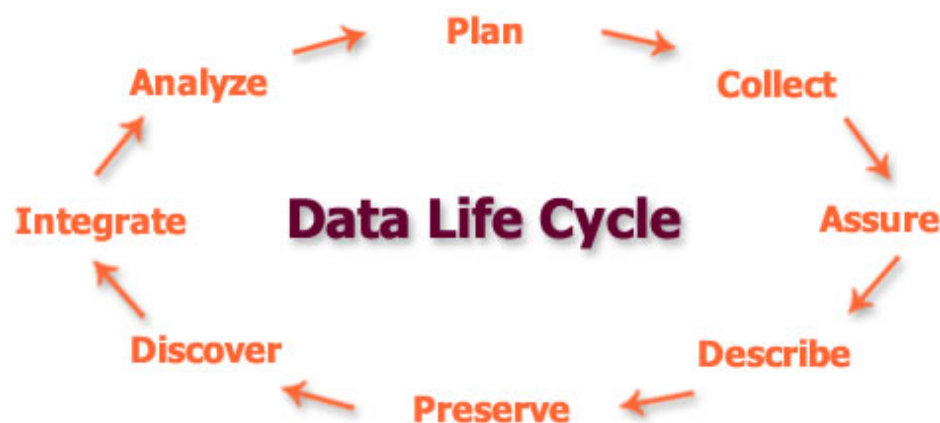
DataONE diagram of the data life cycle.
Credit: DataONE's "Best Practices" [5]

By laying out the blueprint for lifecycle management of data, a DMP provides valuable details, such as how the data will be preserved for the long term, how and where the researcher will make the data available for sharing, and whether reuse of the data, including derivatives, will be allowed.

Reason Two:

Second, related to reproducibility, a DMP can help prevent or reduce the likelihood of mishaps such as data loss, data errors, and unethical uses of data. In effect, a DMP fosters improved communication and accountability for data.
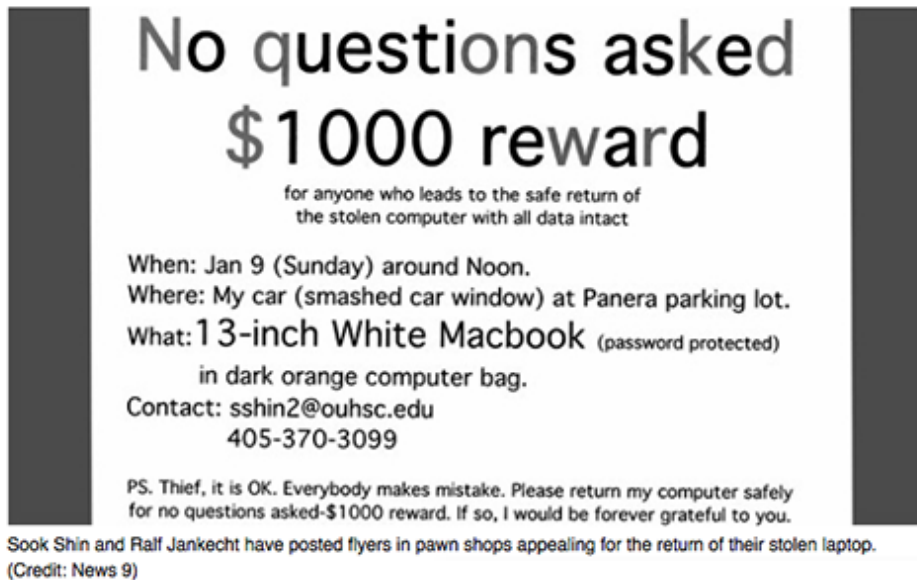
Don't let your research data become a casualty!
Credit: CNET [6]

Reason Three:

Third, data that has been generated by a federally funded project is publicly funded data - that is, data that has been made possible by taxpayer dollars. As such, unless there are restrictions or sensitivities about the data, these are data that should be made available to the public for broad sharing and accessibility.

Researchers in the South Pole gathering ice core samples.
Credit: NOAA [7]

Finally, having a DMP reflects an understanding that the collected data have intrinsic value, as illustrated in the video below. It can be another source of attribution and further investigations. Indeed, as described by Dr. Alfred Traverse, Curator of the Penn State Herbarium, in the following video, sometimes the collected data is all that remains for further investigations.
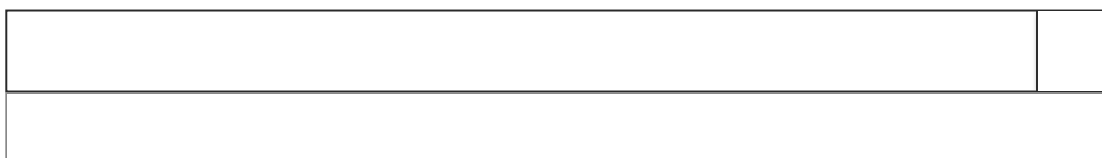
Link to YouTube video. [8]

Components of a Typical Plan

A DMP basically consists of five parts, in which the following aspects of data are addressed:

Part 1: The types of data to be collected or produced during the project, and the processes or methodology for doing so:

- Types of data that will be generated by your research (e.g., human subjects related surveys, field data, samples, model output data)
- Data format(s) and file types (e.g., .txt, .pdf, .xls, .csv, .jpeg, etc.)
- How the data will be collected or accessed (if using existing data)

Part 2: The formats for the data and the standards that will be followed for documenting and describing the data:

- Information about your data you will need to save (i.e., experimental design, environmental conditions, global positioning information, etc.)
- What metadata standard you will use to document your data (i.e., some research domains have widely accepted formats, others may not and you may target how that decision may be made in the project)
- How you plan to record your metadata

CR65 Camponotus floridanus subtracted cDNA library Camponotus floridanus cDNA, mRNA sequence

GenBank: HS411017.1

EST   FASTA

```
LOCUS           HS411017                    182 bp    mRNA    linear    EST 24-AUG-2011
DEFINITION      CR65 Camponotus floridanus subtracted cDNA library Camponotus
                floridanus cDNA, mRNA sequence.
ACCESSION       HS411017
VERSION         HS411017.1  GI:315926471
DBLINK          BioSample: LIBEST_026854
KEYWORDS        EST.
SOURCE          Camponotus floridanus (Florida carpenter ant)
  ORGANISM      Camponotus floridanus
                Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
                Pterygota; Neoptera; Endopterygota; Hymenoptera; Apocrita; Aculeata
                ; Vespoidea; Formicidae; Formicinae; Camponotus.
REFERENCE       1  (bases 1 to 182)
  AUTHORS       Ratzka,C., Liang,C., Dandekar,T., Gross,R. and Feldhaar,H.
  TITLE         Immune response of the ant Camponotus floridanus against pathogens
                and its obligate mutualistic endosymbiont
  JOURNAL       Insect Biochem. Mol. Biol. 41 (8), 529-536 (2011)
  PUBMED        21440063
COMMENT         Contact: Ratzka, C
                Department of Microbiology
                University of Wurzburg
                Am Hubland, 97074 Wyrzburg, Germany
                Email: carolin.ratzka@biozentrum.uni-wuerzburg.de.
```

A Record showing metadata in GenBank.
Credit: National Center for Biotechnology Information [9]

Part 3: The availability of the data, including information about ways in which the data will be accessed, and whether there are any issues related to privacy and/or intellectual property:

- Expected availability of the data during the project period
- List/Explain any ethical or privacy issues incurred by the data
- Address any intellectual property rights issues (e.g., who holds the rights to these data?)

If you are collecting sensitive data (e.g., data stemming from human-subject research), then sharing such data will likely require different types or levels of access.Are higher levels of security required? Will an embargo be needed?
Credit: Patricia Hswe

Part 4: The guidelines, procedures, or policies for data reuse and/or redistribution, attribution, as well as for the creation of derivatives from the data:

- What you will permit in terms of reuse and redistribution of the data, based on policies for access and sharing
- Think about what other researchers (whether in your subject domain or others) may find your data useful
- Identify the lead person or committee on the project who will make the decisions on redistribution on a case-by-case basis
- Where the data will be deposited (e.g., data repository, repository service at your institution, etc.)

Part 5: The measures that will be taken to help ensure the long-term preservation of, and access to, the data - including possible mention of factors such as format migration and who will be responsible for managing the data for the duration of the project:

- Will all of the data produced on your project be preserved, or only some?
- Context for your data (e.g., tools, project documentation, metadata etc.) required to make it accessible and understandable
- Anticipated transformations of the data in order to deposit it and make it available
- The length of time the repository will be available to the public and/or maintained (some directorates have a suggested minimum for the time after a project ends or after publication of certain data)

Again, remember: Data management plans submitted with NSF proposals cannot be longer than *two* pages.

Tools and Other Resources for Data Management Planning

In the years since the NSF and other funding agencies announced the DMP requirement, tools, and other resources have emerged that researchers may find helpful to consult as part of data management planning.

The Penn State University Libraries, in collaboration with the Strategic Interdisciplinary Research Office, have also developed guidance for Penn State researchers that integrates references to the University's research administration policies and guidelines: University Policy Manual [10] and Scholarsphere [11].

Above is an excerpt from the Libraries' Data Management Toolkit.
Credit: PSU Libraries [12]

Information about additional tools, services, and resources for long-term management of data is available from the Libraries' research guide on Data Repository Services and Tools [13].

Another valuable resource is the DMPTool [14], available online for any researcher to use. Penn State has an institutional login [15].
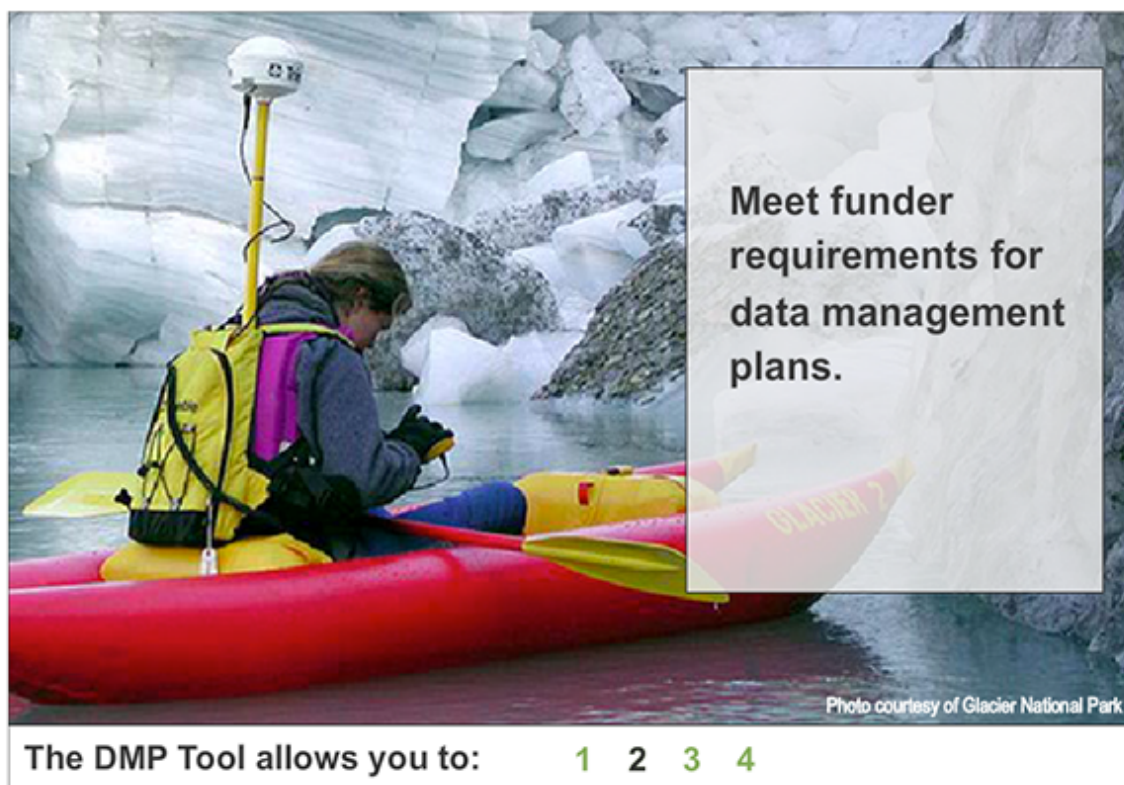
The DMPTool has guidance and templates from various funding agencies and foundations, to help researchers generate a DMP.
Credit: DMPTool [14]

With the DMPTool, researchers complete a webform describing their data management plan, which the DMPTool then formats to the specifications required by the NSF or other major granting agency. The resulting plan, which should be proofed by others (such as the liaison librarian for your subject), will be ready to be submitted, along with the proposal, to the grant funding agency.

If you use the DMPTool to develop a DMP, then keep in mind that the DMP generated at the end might not be only two pages - it could exceed the page limit. This means you'll need to do extra work in making sure the content does not exceed two pages.

Summary

Since 2011, funding agencies such as the NSF, the NIH, and the NEH have required that researchers applying for grant funding for their projects also include a data management plan, also known as a DMP - a document that describes how the applicant will manage the research data that are generated for the duration of the project.

There are many reasons why a DMP is necessary:

1. A DMP gets researchers thinking about the data lifecycle before they start collecting data. It compels them to consider and plan how they will gather, describe, analyze, preserve, and make

accessible and usable their data for other researchers to repurpose or to create derivatives from.

2. Data have intrinsic value that others can learn from and build off of. A DMP helps ensure that data will be available for research verification purposes, if not also reproducibility purposes.
3. A DMP can help stave off data loss and breaches of data, especially for sensitive or restricted data.
4. Research projects funded by federal agencies are projects funded with taxpayer dollars, which means the data should be publicly available. A DMP is intended as additional assurance that such data will be accessible to the public.

Penn State's tools for data management planning include its <u>repository service, ScholarSphere</u> [16]; <u>guidance</u> [11] that integrates Penn State's research administration guidelines and policies for writing a plan; and <u>boilerplate language</u> [17] stating the commitment from the University Libraries and Information Technology Services to preserve and make persistently accessible data sets that are deposited into ScholarSphere. Researchers are welcome to build on the language, in consultation with librarians and technologists at Penn State.

Penn State also has a login for the <u>DMPTool</u> [18], which lets researchers fill in the components of a DMP and, upon completion, then generates a DMP. It is strongly advised that researchers review the resulting DMP to make sure that it does not exceed the two-page limit and that the plan makes sense.

This brief video (4:40) shares a humorous data management and sharing snafu in three short acts:
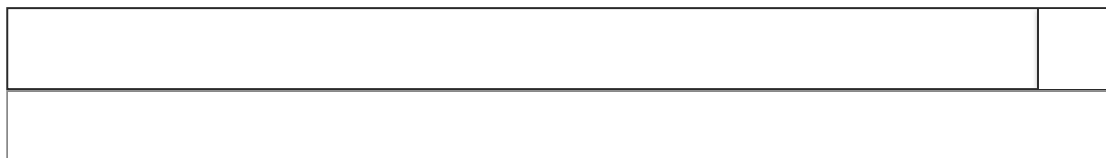
<u>Link to YouTube video.</u> [19]

Click for transcript of Data Sharing and Management Snafu in 3 Short Acts. This will expand to provide more information.

DR. JUDY BENIGN: Hello! My name is Dr. Judy Benign, I'm an oncologist at NYU School of Medicine.

BROWN BEAR: Hello, Dr. Judy Benign!

DR. JUDY BENIGN: I read your article on B-cell function. I think that I could use the data for my work on pancreatic cancer.

BROWN BEAR: I am not an oncologist!

DR. JUDY BENIGN: I know but I think I could use the data for my work on pancreatic cancer. Do you have the data?

BROWN BEAR: Everything you need to know is in the article!

DR. JUDY BENIGN: No. What I need is the data! Will you share your data?

BROWN BEAR: I am not sure that will be possible.

DR. JUDY BENIGN: But your work is in PubMed Central and was funded by NIH.

BROWN BEAR: That is true!

DR. JUDY BENIGN: ... and it was published in Science which requires that you share your data.

BROWN BEAR: I did publish in Science.

DR. JUDY BENIGN: Then I am requesting your data! Can I have a copy of your data?

BROWN BEAR: I am not sure where my data is!

DR. JUDY BENIGN:But surely you saved your data!

BROWN BEAR: I did, I saved it on a USB drive!

DR. JUDY BENIGN: Where is the USB drive?

BROWN BEAR: It is in a box... ... it is in a box at home... I just moved!

DR. JUDY BENIGN: but can I use your data?

BROWN BEAR: There are many boxes! So many boxes! I forgot to label the boxes.

[ON SCREEN TEXT: 7 months later]

DR. JUDY BENIGN: Hello again! Thank you for sending me a copy of your data on a USB drive, I received the envelope yesterday.

BROWN BEAR: You are welcome, but I will need that back when you are finished, that is my only copy!

DR. JUDY BENIGN:I did have a question.

BROWN BEAR: What is your question? You might find the answer in my article!

DR. JUDY BENIGN:No. I received the data, but when I opened it up it was in hexadecimal.

BROWN BEAR: Yes - that is right!

DR. JUDY BENIGN: I cannot read hexadecimal!

BROWN BEAR: You asked for my data and I gave it to you. I have done what you asked.

DR. JUDY BENIGN: But is there a way to read the hexadecimal?

BROWN BEAR: You will need the program that created the hexadecimal file!

DR. JUDY BENIGN: Yes, I will. What is the name of the program?

BROWN BEAR: "Cytosynth"

DR. JUDY BENIGN: I do not know this program.

BROWN BEAR: It was a very good program! The company that made it went bankrupt in 2007!

DR. JUDY BENIGN: Do you have a copy of the program?

BROWN BEAR: I do not use this program any more because the company that made it when a bankrupt. Maybe you can buy a copy on eBay?

[ON SCREEN TEXT: 20 minutes later...]

DR. JUDY BENIGN: I have good news!

BROWN BEAR: You again!

DR. JUDY BENIGN: I talked to my colleague... she knew a person with a copy of the software!

BROWN BEAR: Then why do you need me? Everything you need to know about the data is in the article!

DR. JUDY BENIGN:I opened the data and I could not understand it!

BROWN BEAR: If you have the program you will find it is clear!

DR. JUDY BENIGN: Well... I noticed that you called your data fields "Sam"... Is that an abbreviation?

BROWN BEAR: Yes! It is an abbreviation of my co-author's name... His name is Samuel Lee, we call him "Sam".

DR. JUDY BENIGN: I see... and what is the content of the field called "Sam1"?

BROWN BEAR: Ah yes... "Sam1 is the level of CXCR4 expression.

DR. JUDY BENIGN: and what is the content of the field called "Sam2"?

BROWN BEAR: That is logical if you think about it!

DR. JUDY BENIGN: What is the content of the field called "Sam2"?

BROWN BEAR: I don't remember!

DR. JUDY BENIGN: what about "Sam3"?

DR. JUDY BENIGN: Is there a guide to the data anywhere?

BROWN BEAR: Yes, of course!

It is the article that is published in Science!

DR. JUDY BENIGN: The article does not tell me what the field names mean. Is there any record of what these field names mean?

BROWN BEAR: Yes! My co-author knows what the content of Sam2 is... and Sam3... and Sam4

DR. JUDY BENIGN: Can I talk to your co-author?

BROWN BEAR: I am not sure!

DR. JUDY BENIGN: I would very much like to talk to you co-author.

BROWN BEAR: Well, he was a graduate student. He went back to China 2 years ago.

DR. JUDY BENIGN: Can I have his contact information?

BROWN BEAR: He is in China... his name is "Sam Lee".

DR. JUDY BENIGN: I think I cannot use your data.

BROWN BEAR: You could check the article... to see if what you need is there!

DR. JUDY BENIGN: Please stop talking now!

Check Your Understanding

Why do researchers need a DMP?

(a) It is required by funding agencies.
(b) Having a plan helps ensure data sharing and access.
(c) Replicability of research results depends a lot on good management of data.
(d) All of the above.

Click for answer.

ANSWER: (d) All of the above. The key funding agencies, such as the NSF, NIH, and NEH, are requiring DMPs to foster increased sharing of, and thus access to, research data. Since federal tax dollars fund such projects, the public has a right to have access to the data generated by them. Having a plan for managing data through their lifecycle also aids in the reproducibility of science.

Source URL: https://www.e-education.psu.edu/dmpt/node/644

Links
[1] http://www.youtube.com/watch?v=uHyDzt6E3qU
[2] http://www.nsf.gov/bfa/dias/policy/dmp.jsp
[3] https://www.e-education.psu.edu/dmpt/node/686
[4] http://www.youtube.com/watch?v=5mJ700q6rHs
[5] http://www.dataone.org/best-practices
[6] http://news.cnet.com/8301-17938_105-20028475-1.html
[7] http://www.noaanews.noaa.gov/stories2005/s2419.htm
[8] http://www.youtube.com/watch?v=R2nYe0rq5ac
[9] http://www.ncbi.nlm.nih.gov/nucest/315926471
[10] http://guru.psu.edu/policies/
[11] https://scholarsphere.psu.edu/files/w0892c954
[12] http://www.libraries.psu.edu/psul/researchguides/pubcur/datatoolkit.html
[13] http://www.libraries.psu.edu/psul/researchguides/pubcur/data_resources.html
[14] https://dmptool.org
[15] https://dmptool.org/user_sessions/institution
[16] https://scholarsphere.psu.edu/
[17] https://scholarsphere.psu.edu/files/cv43nw822
[18] http://https://dmptool.org
[19] http://www.youtube.com/watch?v=N2zK3sAtr-4