Guide for Creating a PEER Data Management Plan

Table of Contents

Table of Contents 0

Overview 1

Getting Ready To Write The DMP 2

- **Ask your US partner and colleagues** 2
- **Find a data repository** 2
- **Legal restrictions** 2

Template: Data Management Plan 3

Instructions For Data Management Plan Questions 4

- **Types of Data Produced** 5
- **Metadata** 8
- **Data storage and preservation** 10
- **Policies for access and sharing** 11

Overview

A Data Management Plan (DMP) is a document that describes how you will manage your data during the research project and what happens to the data after the project ends. A DMP is a plan for your data and so it should be written before a project begins, however it is never too late to write one if the project is already underway. It will help you to think ahead to the journal publications you will write based on the PEER award and to plan for how you will organize and then share the data that you will use for the publication.

A comprehensive DMP will discuss the following aspects of the data life cycle:

- Collect - How the data is collected and processed by the researcher.
- Assure - How to make sure the data is high quality and free of errors.
- Describe - How the data will be documented so that other researchers can use it.
- Preserve - How and where the data will be stored so that researchers can access it forever.

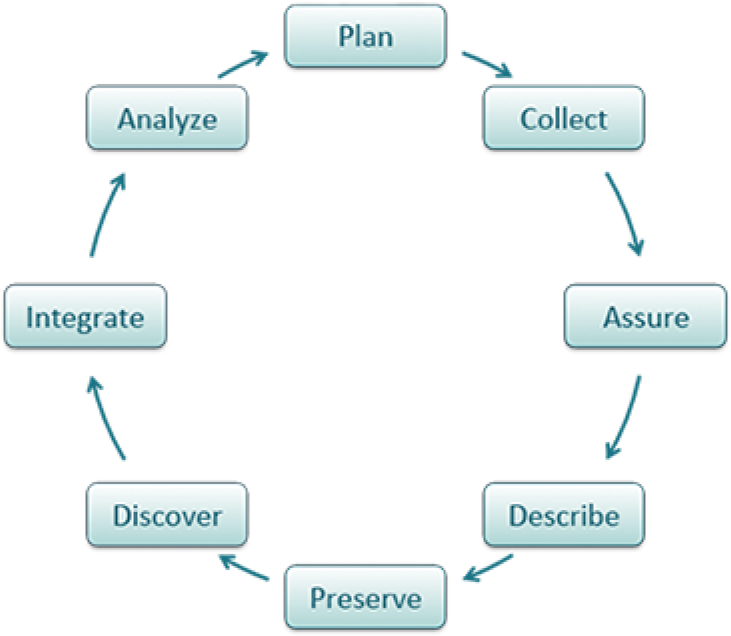

Figure 1: The Data Life Cycle. Credit: DataONE

Getting Ready To Write The DMP

Before you begin your own DMP, there are four things you should do first.

Ask your US partner and colleagues

Ask your US partner and colleagues if they have a data management plan! It is possible that they have already thought through most of the issues that you might have. If they have a DMP, then you should definitely request a copy. It will be a useful guide for you.

Find a data repository

Look through data repositories that are relevant to your research area. Search for data that is similar to yours. Search for data that you want to have. Make a note of the data repositories you think would be a good place to upload your data.

If you do not already know which data repositories are popular in your research area, then you should ask your US Partner for their advice. Otherwise, Nature has a short list of its recommended data repositories for each research area. Many PEER projects work with Earth and environmental data, in which case Pangaea is a good place to start. If you are doing social science, health, or agricultural research, you can store your data for free in USAID's data library, the DDL. You can also use this search engine to look through many different data repositories at once; search for data like yours and see which repositories come up most often. If you cannot find a repository that seems appropriate for your work, you can submit your data to The Harvard Dataverse, which freely accepts data of all kinds.

Legal restrictions

Are there any reasons why you might not be able to legally share your data? Are you aware of any restrictions on sharing data that apply to your country? Are you using data or tools from private sources that may not be able to be legally shared? Ask the lawyers at your university. Ask the appropriate Ministry. Ask your colleagues. Open data makes for strong science, but it may not be legal in all cases. You should begin these conversations immediately because it can take time to find answers.

Template: Data Management Plan

In the first year of your PEER award you are asked to create a data management plan. You will create this plan in Foundant by completing the Data Management Plan report which will be assigned to you by your grant manager. The questions you will complete are listed below.

Types of Data Produced:

1. *What data will be generated in the research?*
2. *Are any of the data that you plan to work with considered sensitive? If so, how will you safely handle it?*
3. *What is the estimated size of the data set that you will generate?*
4. *Will you be using existing data? If so, how you will obtain it?*
5. *How will the data be processed?*

6. *What quality control methods will you apply to your data?*
7. *What software tools will you use to organize and analyze your data?*

Metadata:

1. *How will your data and analysis be documented?*
2. *Which metadata standards will you use and why have you chosen them?*
3. *What details about your data, workflow, and analysis will you document?*
4. *Which file formats will you use for your data when you upload it to a repository?*

Data storage and preservation:

1. *Short-term data storage: how will you backup your data?*
2. *Long-term data preservation: which data repository will you use?*
3. *When will you upload your data to a repository?*

Policies for access and sharing:

1. *Who is responsible for managing the data?*
2. *Will you request an embargo for your data? If so, why?*
3. *Are there ethical and privacy issues with sharing your data? If so, how will these be resolved?*
4. *What license will you apply to your data?*

Instructions For Data Management Plan Questions

Each funding agency will have its own requirements for what should be in the plan. When writing a DMP for any of your other grant proposals, be sure to check what requirements apply for that specific grant. Luckily, the requirements at all funding agencies are very similar; so writing this DMP for your PEER award will prepare you for writing future DMPs.

You will find the DMP template in Foundant. You will write and submit the plan through Foundant. The DMP has four main sections and each section has a list of questions *in italics* to be answered. Please answer every question. If a question does not apply to you, then just write "N/A".

Each question from the template is reprinted below, along with an explanation and sample responses that are drawn from the DMPs of other researchers. While this may seem like a lot of questions to answer, the responses to most questions can be short. A DMP is often less than two pages (with the questions deleted).

Types of Data Produced

What data will be generated in the research?

A single project may generate many different types of data. Examples of research "data" include observational data, physical samples, results from models and simulations, human behavior survey results, software or algorithms, etc. Please **briefly** describe the different types of data that you expect to generate.

Sample:

- *This project will use trained bird watchers to count the occurrence of ten songbird species and one red squirrel species in high elevation forests of the northeastern United States and southeastern Canada.*
- *The program captures wave, wind, and temperature data in real-time, updating every 30 minutes.*

Are any of the data that you plan to work with considered sensitive? If so, how will you safely handle it?

This question is concerned with keeping the data safe *while you are collecting and analyzing it*. Later in the DMP you will answer a similar question about how you will make the data safe *to share*.

Please answer yes or no. If yes, what procedures will you use to protect the data while it is *in your possession*?

Sample:

- *Yes. The data will be processed and managed in a secure non-networked environment using virtual desktop technology.*
- *Yes. The data files from this study will be managed, processed, and stored in a secure environment (e.g., lockable computer systems with passwords, firewall system in place, power surge protection, virus/malicious intruder protection) and by controlling access to digital files with encryption and/or password protection. De-identified files will be deposited with [repository] whose security policy has been written according to best practices.*
- *No.*

What is the estimated size of the data set that you will generate?

It is helpful to know how large your data set will be when the experiment is complete. Some data sets are very small when stored on a computer (less than 1 megabyte) which makes them easy to manage and eventually upload to a repository. However, larger data sets can be more difficult to manage and to upload. If you know that you will have many gigabytes of data at the end of the project, that can change the data repository that you will use to upload your data.

Sample:

- *The stream gauges in the field generate about 0.5 MB of data per day in total. They will collect data for two years, so we expect to generate 0.5 MB * 2 years * 365 days = 365 MB of data.*

- *Stored in a comma separated spreadsheet file, each survey is about 1 KB of data. We will administer 600 surveys. Thus we will collect about 0.6 KB of data.*

Will you be using existing data? If so, how you will obtain it?

A research project may generate new data, analyze previously generated data, or do both. If you will use any *previously generated data* that has been collected for a different project or by someone else, then you should describe what the data is and how you will obtain it.

Sample:

- *We will use medical records gathered from 32 health care facilities in the target district. These records provide data on malaria exposure and treatment of mothers and their children. To obtain the data, we have an agreement with the Ministry of Health to access the national health data registry.*

How will the data be processed?

Please describe any actions that will need to be taken on the raw data in order to prepare them for analysis.

Sample:

- *The medical records are paper-based and will be digitized. The digitization will be done by a contractor.*
- *These measurements will be recorded into a lab notebook, then entered into an Excel spreadsheet. At each sampling location we will record the coordinates. These coordinates will be used to create a map displaying sampling locations. ESRI's ArcMap program will be used to create the map*

What quality control methods will you apply to your data?

Everyone makes mistakes, but what will you do to find and correct the mistakes you make? High quality data is essential to good science and every effort should be taken to improve the quality of your data. Quality control methods may include:

- how to identify potential errors in the data;
- how to correct those errors;
- how problematic data that cannot be fixed will be marked.

For further information, please see this resource on how to ensure basic quality control.

Sample:

- *Quality assurance measures will comply with the standards, guidelines, and procedures established by the World Health Organization.*
- *For quantitative data files, the repository ensures that missing data codes are defined, that actual data values fall within the range of expected values and that the data are free from undefined codes. Processed data files are reviewed by a supervisory staff member before release.*

What software tools will you use to organize and analyze your data?

Small amounts of data can be managed by programs like Microsoft Excel. Larger or more complex data sets may require a database tool (mySQL) or a GIS tool (ArcGIS). If you are producing a software program as part of your research, this is also considered 'data'; you are *highly encouraged* to use a form of 'version control' to manage your software development process (examples: git or svn). Please describe the tools that you will use to organize your data.

Sample:

- *All data will be entered into a MySQL database. The data will be analyzed with scripts that are written in Python.*
- *Observational data will be entered as rows in an Microsoft Excel spreadsheet, which will be exported to a database for use with ArcGIS.*

Metadata

How will your data and analysis be documented?

Good documentation is required in order for other researchers to understand and use your data.

- The simplest way to document your data is to have a plain text "README" file in the main folder that holds your data. This file would describe the contents of the folder and information regarding how the data was collected.
- A more sophisticated method would be to prepare a full document that rigorously describes all components of the data set (e.g. codebook). It is possible that your discipline provides software tools to help you manage your metadata and to produce such a document.
- The best way to document your data is to produce a "data paper". This is a peer-reviewed journal article that discusses all aspects of how the data was collected and processed, however it does not contain any analysis, interpretations, or conclusions about the data. Along with the "data paper", you would upload the data. This is the best method for three reasons. 1.) It provides the most complete documentation of your data, 2.) it is very easy for other researchers to cite your data when they use it because they can cite a journal article, and 3.) it provides you with a peer-reviewed journal publication.

Please describe the method that you would like to use to document your data. Further, discuss the information that you will include in the documentation.

Sample:

- *I will include a README file that follows the template provided by Cornell at https://cornell.app.box.com/v/ReadmeTemplate.*
- *The documentation will be generated by the tool Morpho and distributed as a pdf document.*

Which metadata standards will you use and why have you chosen them?

For many types of data, there are standards / guidelines for what metadata to use. To choose a metadata standard, it is helpful to consider the data repositories where you would like to upload your data. Many repositories recommend metadata standards. Some recommended tools are:

- DDI - for Social, Behavioral, and Economic Sciences - Nesstar Publisher (Win)
- EML - for Earth, Environmental, and Ecological Sciences - Morpho (All)

To answer this question, do one of the following:

- Choose a metadata standard that is popular in your field and say why you chose it; or
- Explain that you were unable to find an appropriate standard.

Sample:

- *The biological and ecological data will be structured in Ecological Metadata Language (EML).*
- *All physical and chemical time series data will be formatted to follow the standard operating procedures for ocean acidification research as described by Riebesell et al. (2010)*

What details about your data, workflow, and analysis will you document?

The data is important, but it is just as important to document how the data is created and processed. In this section, list the details you will document about your research process. For instance: If you're using lab equipment to process your samples, will you list what the equipment is? If there is calibration data for the equipment, will you provide it? A few years from now, one of the common pieces of equipment in your field might be shown to have a hidden flaw; it will be very helpful to know if you used that machine or not. If you're working with observational data, how will the observers be trained? Different birdwatchers might have different biases or skills depending on their training. What hardware and software will you use to process the data? Will you document any computer source code that you develop? Etc.

Sample:

- *The documentation will contain the survey questions distributed to villagers, a description of how the survey was administered, the codes corresponding to the possible responses, and statistics for how each of the questions were answered. R code will be annotated for those seeking to replicate the analysis.*

Which file formats will you use for your data when you upload it to a repository?

A 'file format' is a standard way that information is stored in a computer file. For example, most pictures are stored as a JPEG / JPG or GIF. The file format for this training document is PDF. It is important to think about the formats in which you will store your data, because many file formats are proprietary / private. For instance, if you store your data in a Microsoft Excel spreadsheet and save it as an XLS file, then only other people who have paid for Microsoft Excel can access your data. Use of proprietary formats is highly discouraged, with a few exceptions.

Sample:

- *The survey data will be distributed in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, and Stata.*
- *Digital image data will be processed and submitted to the repository in TIFF version 6 uncompressed (.tif) format.*
- *Geospatial data will be processed and submitted to the repository as an ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn).*
- *Textual data will be processed and submitted to the repository as plain text data, ASCII (.txt).*

Data storage and preservation

Short-term data storage: how will you backup your data?

Please describe how your data will be stored while you are collecting it and working with it. Perhaps you will fill many notebooks full of handwritten data, but what happens to that data if there is a flood or a fire? If you will store all of your data in a spreadsheet on your computer, what happens to that data if your hard drive fails? Every year, there is a 2% chance that your hard drive will fail. It's best not to take chances. It is highly recommended that you "backup" your data.

If you are generating only a small amount of data, less than 2 GB, you can store this easily, securely, and for free on Dropbox. If you will generate more than this, you can store up to 1 TB of data securely and easily on Dropbox for a fee of $100 per year. This might seem like a lot of money, but it will also ensure that you don't lose years of your work due to an accident. If you would prefer not to use Dropbox, here is a list of 10 alternatives. It is also possible that your university offers a data backup service; you should ask your university about this.

Sample:

- *Data files will be kept in a folder that is automatically backed up by Dropbox. The data does not contain any sensitive information, so there are no security issues with using Dropbox for this purpose.*
- *Due to the sensitive nature of our data, the computer on which the data is stored cannot be connected to a network. Thus, we are unable to backup our data on a remote server.*

Long-term data preservation: which data repository will you use?

Please provide a link to the data repository that you plan to use and explain why you chose it. See the Open Data Pilot Program overview document for help. Keep in mind that some repositories require an agreement or fees. Your choice is not permanent and you are free to choose a different repository later.

Sample:

- *The majority of the data will be available on the open access repository Pangaea (www.pangaea.de).*

When will you upload your data to a repository?

PEER will require that data be uploaded within 30 days of a journal article being accepted for publication. However, it is encouraged that the data be made available at the same time as the journal article so that the data can be cited within the article. There might also be a reason why it will take you longer than 30 days to upload your data after your publication. Please describe when you plan to upload your data.

Sample:

- *Data will be uploaded to the repository once the journal article has been accepted for publication.*

Policies for access and sharing

Who is responsible for managing the data?

This person will be the one who is responsible for controlling access to the data and making sure that the data is properly documented and shared. They are in charge of the short term storage and long term data preservation. You must name the individual.

Sample:

- *Thomas [censored], a post-doc in the PI's lab, is responsible for managing the data.*

Will you request an embargo for your data? If so, why?

If you would like to request an embargo, please explain why it is necessary and give the length of time the data would be embargoed. If you do not need an embargo, please say so.

Sample:

- *A one-year embargo is requested because the co-PI is currently drafting a journal article on the same data set.*
- *The data will not be embargoed.*

Are there ethical and privacy issues with sharing your data? If so, how will these be resolved?

Please refer to the Open Data Pilot Program overview document for a discussion of sensitive data and data anonymization.

Sample:

- *The results of this study will not contain any personal information of the participants involved, nor will it contain any unique personal identifiers that could cause assumptions to be made about those involved. In order to assure that the results can be shared publicly, precautions have been taken to ensure that the participant's identities are kept anonymous.*
- *The data set will be anonymized by removing all of the HIPAA Safe Harbor elements. Further, we will follow the data anonymization procedure prescribed by the Ministry of Health.*

What license will you apply to your data?

Will you use Attribution, Attribution-ShareAlike, or Public Domain?
Sample:

- *The data will be released into the Public Domain under the ODC-PPDL license.*

Background: When you upload your data to a repository, you must also give other researchers the *legal rights* to use the data. This is done by choosing a "license" to apply to your data set. A license is legal language that describes the ways you wish to allow your data to be used. Most licenses will require that the original creator of the data be cited if the data is used. The license can either allow or forbid commercial uses of the data. The license can allow others to modify the data or forbid them from making any changes to the data set. It is up to you which license you prefer to choose. Be aware that some data repositories will only accept data if it uses their preferred license.

There are two main organizations that write the language for the licenses: Creative Commons (CC) and the Open Data Commons (ODC). There are three types of license that apply to research data, which both CC and ODC provide. The license types are:

- Attribution
    - This license lets others distribute, change, and build upon your work, even commercially, as long as they credit you for the original creation.
    - The CC-BY and ODC-By licenses are this type.
- Attribution - Share Alike
    - You let others copy, distribute, display, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first. This is the license used by Wikipedia.
    - The CC-BY-SA and ODC-ODbL licenses are this type.
- Public Domain
    - All rights worldwide to this work are waived by the author. Others may copy, modify, and distribute the work, even for commercial purposes, all without asking permission.
    - The CC0 and ODC-PPDL licenses are this type.

The Cornell University website hosts a good discussion of data rights and contains the following explanation for why you should consider using a Public Domain license:

> There is no single right answer as to which license to assign to a database or content. Note, however, that anything other than an ODC PDDL or CC0 license may cause serious problems for subsequent scientists and other users. This is because of the problem of attribution stacking. It may be possible to extract data from a data set, use it in a research project, and still maintain information as to the source of that data. It is possible to create a data set derived from hundreds of sources with each source requiring acknowledgement. Furthermore, the data in the other databases may not have originated with it, but instead sourced from other databases that also demand attribution. Rather than legally require that everyone provide attribution to the data, it might be enough to have a community norm that says "if you make extensive use of data from this data set, please credit the authors."

Sample:

- *The data will be released into the Public Domain under the ODC-PPDL license.*