PURDUE UNIVERSITY | Libraries

What Should Your Data Management Plan Address?

A Guide for Purdue University Researchers

This guide was created by Purdue University Libraries in accordance with the National Science Foundation (NSF) data management planning requirements. This information is also available as a DMP template on DMPTool.org. Instructions for accessing the DMPTool template, as well as other data management resources, are available on the Purdue University Research Repository (PURR) website.

YOUR DMP SHOULD INCLUDE

1. **Data Types**
   The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
2. **Metadata**
   The standards to be used for data and metadata format and content
3. **Access and Sharing**
   Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
4. **Re-use**
   Policies and provisions for re-use, re-distribution, and the production of derivatives
5. **Archiving**
   Plans for archiving data, samples, and other research products, and for preservation of access to them

1 Data Types

1.1 Conducting a Data Inventory of Your Project

Research data are formally defined as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings" by the U.S. Office of Management and Budget (https://science.energy.gov/funding-opportunities/digital-data-management/).

As data can take many forms and can actually change forms during the course of a project, it is important to first create an inventory of potential data to be generated in order to then plan for managing that data.

Possible Questions to Consider

What types of research data will your project generate?

One project may generate a number of different types and forms of data. Besides simple text files that an analytical instrument might generate, other forms of data might be field notes, survey data, image files, lab notebooks, software created for the project, curriculum materials, and other similar output of your research that will need to be managed throughout the course of the project.

What stages might your data go through?

Again, a text file from a direct reading of an instrument might be categorized as “raw” data, but you might also use software to re-format the text file for analysis, thus creating a “processed” data set. Further, this “processed” data might be used to create graphs or equations, leading to an “analyzed” data set. Data sets that are publication ready might be considered “finalized.” It is useful to think of the life cycle your data sets might take throughout the course of the project to help you design an effective data management plan.

What will your data “look” like? What are the “technical specs” of your data?

Consider the format of the data sets—file types, average file size, volume, and/or estimated number of data files produced. How quickly will the data be generated—all at once or over the course of a number of years? If non-electronic data will be collected, consider whether that data will be converted to a digital format or remain analog.

What documentation will be generated for this data?

In order to be reproducible, some documentation will need to be associated with the research data generated by a project. Will the documentation be of sufficient depth and quality for others to be able to understand and make use of this data? Just as important a consideration is who will be responsible for the documentation.

Resources

The DCC Curation Lifecycle Model, from the Digital Curation Centre
http://www.dcc.ac.uk/resources/curation-lifecycle-model

Conceptualizing the Digital Life Cycle, from the International Association for Social Science Information Services and Technology
http://www.iassistdata.org/blog/conceptualizing-digital-life-cycle

1.2 Selection and Appraisal Policies

Not all data will need to be kept forever, and after manipulation and processing, it may only be necessary to keep the "raw" and "final" versions of the data. Making decisions on what data sets to keep and subsequently share is very important as multiple versions of data are more difficult and expensive to migrate, preserve, and provide access compared to only selecting the core data for inclusion within a data management plan.

Possible Questions to Consider

What is the value of your project's data to others?

As you begin to think about the various types of research data generated by your project, consider what components of your data might be valuable, both in general and for specific audiences, in terms of re-use or in combination with existing data.

What is the possible size of the user community that might be interested in your data?

Will only those in a narrow discipline be interested or does your data have broader, interdisciplinary appeal? Is your data of local, national, or international interest?

How long would this data be valuable to others?

Consider the nature of your discipline or field of research and the longevity of the usefulness of your data—5 years, 15 years, 50 years, indefinitely, etc.

How unique is your data?

Are you carrying out research that is similar to other research that is publicly available or are you creating a comprehensive data set on a particular subject that has never been collected before? Is your data complete enough to be useful on its own or will it be better utilized within the context of other data that is available?

Resources

Appraise and Select Research Data for Curation, from the Digital Curation Centre
http://www.dcc.ac.uk/resources/how-guides/appraise-select-research-data

Appraisal and Selection, a chapter from the Curation Reference Manual, from the Digital Curation Centre
http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/appraisal-and-selection

The Selection, Appraisal and Retention of Digital Scientific Data, final report of the 2003 ERPANET/CODATA Workshop
http://www.erpanet.org/events/2003/lisbon/LisbonReportFinal.pdf

2 METADATA

2.1 DATA STANDARDS

Data standards are typically the file format in which your data is generated. (Examples of standard data formats include: XML, ASCII, CSV, MySQL, netCDF, etc.)

Agreeing to particular data standards will not only allow you to share data with all project personnel, but will also allow others to view or use your data once it is shared. The more common the data standard, the more universal use and acceptance your data is likely to have. Using a standard data format will also decrease the likelihood of your data becoming obsolete because it can no longer be read and will help others archive your data for long-term preservation if necessary.

Conforming to data standards might be as simple as making sure your project personnel are not saving spreadsheets from three different versions of Microsoft Excel. It might also mean having to convert data taken from one piece of software and converting it to another data format so it can be easily shared with others (such as converting a word processing file into a more standardized ASCIII or XML data standard).

Possible Questions to Consider

Will your data conform to any standard formats already agreed upon by those in your field or discipline?
Whether you will be using them or not, consider what data standards are currently accepted within your field. It will be easier to compare and share data with others if your data is similar to others' in your discipline.

How accessible are your chosen data standards for others to read and/or re-use your data?
Consider whether the data standards you have chosen are Open Source or may require specific software or instrument to read it.

Who on the project will be responsible for properly applying these data standards?
Those responsible should ensure the agreed upon data standards are properly and consistently applied. Following-up to make sure the data is properly formatted for later dissemination and re-use by others is also an important factor.

Resources

Data Formats Table (data formats currently recommended by the UK Data Archive for long-term preservation of research data)
http://www.data-archive.ac.uk/create-manage/format/formats-table

Open format - listing of open (non-proprietary) file formats from Wikipedia
http://en.wikipedia.org/wiki/Open_format

2.3 Metadata Standards

Metadata can be defined as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (such as a data set).

"A metadata record is a file of information which captures the basic characteristics of a data or information resource. It represents the *who*, *what*, *when*, *where*, *why*, and *how* of the resource."

Therefore, properly associating metadata to the research data you generate will allow yourself and others to efficiently locate, share, and re-use your data since each data set will be properly documented. Providing metadata for your data sets also asserts your authorship of the data and allows others to credit you when necessary.

Some metadata standards are very general and can be applied to a variety of situations, others are more discipline-specific. Examples of general metadata standards include Dublin Core (DC), Resource Description Format (RDF); and specific ones include: Federal Geographic Data Committee (FGDC), Directory Interchange Format (DIF), Ecological Metadata Language (EML), Proteomics Metadata Interchange Schema (PROMIS), and the Data Documentation Initiative (DDI).

Possible Questions to Consider

How do you expect your data sets to be described and documented?

Will there be detailed annotations, a code book, a data dictionary, or similar system of describing your data? Are your descriptions based on an internal system or a more universal standardized system of description? Thinking ahead, can this system help your project conform to a standardized metadata standard?

Are there any standard formats already agreed upon by those in your field or discipline that would be appropriate for your project to use?

Some disciplines have specific metadata standards developed within the discipline, while for other disciplines a general metadata schema would be just as acceptable. If you read the literature in your field, do others talk about particular metadata standards or similar systems of describing their data?

Who on the project will be responsible for properly applying these metadata standards?

Those responsible should ensure the agreed upon metadata standards are properly and consistently applied. This will ensure that others will be able to find, understand, and make use of your data once it has been disseminated.

Resources

Glossary of Metadata Standards
http://jennriley.com/metadatamap/

Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata
http://dublincore.org/documents/dc-citation-guidelines/

Directory Interchange Format (DIF) Writer's Guide (used by NASA’s Global Change Master Directory)
http://gcmd.gsfc.nasa.gov/User/difguide/difman.html

2.4 File and Directory Naming Conventions

Because there can be multiple sources of data for a research project, multiple researchers working on the same data files from a project, and turnover in project personnel, it is important to develop a standardized method for naming your data files and directories.

Possible Questions to Consider

How will you name your files and create your directory structure?

In doing so, will they be able to be sorted in a logical manner (such as by date or trial run number)? Can file and directory names be interpreted without a complicated translation key? Always maintain the original file extension generated by the source program/instrument. Even if they are compatible, for future use, a changed file extension can be misleading (such as renaming a *.docx file to a *.doc file without actually saving it in the new file format).

How will you handle multiple versions of the same data?

File version control is important, especially since only minor changes may be made on an infrequent basis. Consider how you will keep track not only of the original data file in case subsequent manipulation is in error, but also how you can easily identify what is the latest version of a particular data file.

How will your naming conventions be documented and who will be in charge of maintaining their consistency?

Since project personnel may change and for future dissemination of your data, it is important to document how your file and directory names are being generated, especially if you are using particular abbreviations for instruments, techniques, geographic locations, test subjects, etc.

Resources
Best Practices for File-Naming
http://digitalpreservation.ncdcr.gov/filenaming.pdf

3 Access & Sharing

3.1 Components of Access & Sharing policies

Sharing research data is now a requirement of the National Science Foundation and other funding agencies. However, there are many issues to consider and decisions to make surrounding making your data available to others. These decisions should be clearly addressed in written policies that are included as components of an overall data management plan. They should also be addressed as a part of a submitted proposal.

Possible Questions to Consider

What data will you make available to others?

Your proposed project may produce multiple sets of data, or pass through different stages (raw, processed, analyzed, etc.). You may decide that not all of your data may be suitable for open access to the public. If you decide not to share a particular data set, you should explain your reasoning in your proposal.

How will you make your data available?

Is there a repository for the type of data that you will generate that could host your data and make it available to others? If not, you will need to consider how you will ensure stable and ongoing access to your data. Purdue has developed this site as a resource for such a purpose.

When will you make your data available?

The NSF Award and Administration Guide provides some flexibility on when data should be made available, stating that data should be shared "within a reasonable time". A well-developed data management plan would list when the data would be made available to others and provide a justification for doing so.

> **Example:** "The data will be shared with others immediately after the findings derived from this data have been published, to give project researchers the opportunity to benefit directly from their work."

Any restrictions on access, such as an embargo or limiting access to specific groups, should be explained in the sharing and access policy. A description of how data will be shared should include information about access procedures, embargo periods, the technical mechanisms for dissemination, and whether access will be open or granted only to specific user groups. A timeframe for data sharing and publishing should also be provided.

Who are the intended audiences for your data?

The NSF expects that data from funded projects will be shared with other researchers. However identifying the research communities or other populations who will likely find value or make use of your data is good practice, and will help guide decisions on use of standards, descriptive metadata, and other usability issues.

Resources
ICPSR Access Policy Framework
https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/access-policy-framework.html

3.2 Privacy and Confidentiality / policies

Some data sets are generated from human subjects and may contain sensitive information that they would not wish to be revealed publicly. Researchers at Purdue are already bound to follow the policies and procedures of the Human Research Protection Program, however sharing the data itself beyond the research team will assuredly require additional thought and planning. Researchers should be sure to consult the Human Research Protection Program office for advice and guidance in crafting a data management plan that adheres to proper protocols and procedures.

Beyond the use of Human Subjects, other types of research requiring institutional review may also need to fulfill additional requirements before the data can be made accessible to others. Careful thought and consideration should be given to the information that will be either directly or indirectly conveyed by the data upon its release, and whether or not this information warrants the development of measures for protection.

In addition to Purdue's Human Subject Protection Program, a number of professional societies have produced codes of ethics. When developing data management plans for a proposal, researchers should consult relevant disciplinary code of ethics and follow their prescribed best practices.

The NIH suggests the following strategies for addressing privacy and confidentiality issues: "withholding part of the data, statistically altering the data in ways that will not compromise secondary analyses, requiring researchers who seek data to commit to protect privacy and confidentiality, and providing data access in a controlled site (a data enclave)."

Possible Questions to Consider

How will you protect the identity of the human subjects from being disclosed in making your data accessible to others?

In considering this question, attention should be given not only to removing information that directly identifies human subjects, but to addressing any indirect information that could be pieced together and used to identify subjects as well.

How will you demonstrate compliance with federal, professional, institutional, and any other regulations or guides that cover the research conducted and the resulting data sets?

Resources
The American Statistics Association, "Ethical guidelines for statistical practice." (in particular: Section D "Responsibilities to Research Subjects"
http://www.amstat.org/about/ethicalguidelines.cfm

Purdue's Human Research Protection Program
https://www.irb.purdue.edu/

3.3 Security issues / policies

A well-developed data management plan will address security from multiple viewpoints. From a broad perspective, data security issues include the technological infrastructure used to store and disseminate the data, as well as the physical environment in which this infrastructure and data reside. At a more granular level data security includes ensuring the data are reliably backed up, protecting the data (and the physical infrastructure that host the data) from unauthorized access or unwanted alterations.

It is important to keep in mind that managing different data types may require different levels of security. For example, data that contain confidential or sensitive information may require additional security measures than data without such information. One way to determine the level of security needed for a particular data set would be to determine what the consequences of accidental disclosure or loss of the data would likely be.

Possible Questions to Consider

What actions will be taken to ensure the security of the data?

A good data management plan will identify security issues surrounding the data and describe specific actions that will be taken to address these security issues.

Who will have designated role(s) in ensuring the security of the data? What specific responsibilities will be assigned to these roles?

Although the Principle Investigator of the project will likely have the ultimate responsibility for planning, implementing and ensuring the security of the data, the individual roles and particular responsibilities of project personnel should be identified in a data management plan.

Will the data require different or additional security measures once they are made accessible to others?

Security needs and considerations may change once the data are made available outside of project personnel. Security needs related to sharing data should be considered and developed in advance of releasing the data.

For consultation on security issues in sharing and access research data, contact Secure Purdue here.

For help in finding information related to security issues in sharing research data, contact a Purdue Libraries' faculty member for your subject area. A list of faculty librarians is available here.

Resources

Purdue's Information Security and Privacy policy (although not geared towards research data specifically, this document still provides an example of a data security policy)
http://www.purdue.edu/policies/information-technology/viib8.html

The UK Data Archive: Data Security
http://www.data-archive.ac.uk/create-manage/storage/security

3.4 Intellectual Property & Copyright issues / policies

Sharing data effectively requires a careful consideration of copyright and intellectual property issues. Although facts and common knowledge cannot be protected under copyright law, the arrangement, interpretation and expression of data are protected. Data can be licensed, giving researchers the ability to set conditions on how the data may be used by others. Although funding agencies generally acknowledge intellectual property rights for data and other outputs of funded projects, increasingly there is an expectation that investigators will share their data with others within a reasonable time and at a minimal cost. See Section D – Intellectual Property in the NSF's Award and Administration Guide as an example.

Possible Questions to Consider

Who owns the data?

Issues of ownership will affect the terms and conditions under which the data are administered, shared, and preserved. Questions of ownership, and perhaps more importantly defining who has what responsibilities and authority over the data, should be discussed and settled early on in the project.

Will responsibility for the data be transferred to another party (such as a data repository) for making the data publicly available, preservation purposes, or other reasons?

If so, when and how will the transfer of data take place? What arrangements and agreements need to be crafted to support the transfer of data?

How will you license your data?

Issues to consider include: attribution (see section 4.2 below), conditions on re-use of data by others, conditions on the redistribution of data by others, conditions on the creation and publication of derivatives of the data, and conditions on using the data for commercial purposes.

For assistance with copyright and intellectual property Issues, contact Donna Ferullo at the University Copyright Office.

Resources

Creative Commons - "Creative Commons provides free licenses and other legal tools to mark creative work with the freedom the creator wants it to carry, so others can share, remix, use commercially, or any combination thereof."
http://creativecommons.org/

Guide to Open Data Licensing - "A guide to licensing data aimed particularly at those who want to make their data open."
http://www.opendefinition.org/guide/data/

4 RE-USE

4.1 INTELLECTUAL PROPERTY AND REUSE

Funding agencies may have varying approaches towards intellectual property, copyright and related issues. Please check with your funding agency and program officer(s) with any questions about specific requirements or questions about your data. Intellectual property rights for data sets are subject to University policies. See http://www.lib.purdue.edu/uco/Resources/campus.html for a list of campus resources on copyright and intellectual property.

Possible questions to consider

Who will own these data sets? Any other stakeholders need to be consulted before data sets are made available?

Will you permit the re-use of the data, either with or without conditions?

Will you permit the re-distribution of the data, either with or without conditions?

Will you permit the creation and publication of derivatives from the data, either with or without conditions?

Will you permit others to use the data to develop commercial products or in ways that produce a financial benefit for themselves, either with or without conditions?

How will the people who generated the data sets receive attribution for their work?

4.2 ATTRIBUTION/ACKNOWLEDGEMENT/CITATION

Scholarly research has been producing increasing amounts of data, and has been increasingly reliant on data to develop and share research results. In this context, appropriate attribution of research data is critical for:

- Providing systematic and persistent access to data sets
- Increasing the acceptance of data as a legitimate part of the scholarly record
- Supporting data archiving that will permit results to be verified and re-purposed for future study

Components of a data citation

At this point, standards for citing data sets are not fully agreed upon. Some, but not all, style guides have recommendations for citing data. APA, for example, includes such recommendations. Some data providers, such as ICPSR, also have specific guidelines for citing their data. Despite this lack of a formal standard at this point, data citations should include the following elements:

- Author(s)
- Title
- Year of Publication
- Publisher (Often the data archive housing the data set)

- Version
- Access Information

The access information element may consist of a URL and/or a persistent identifier such as a DOI. Purdue University is able to assign DOIs to data sets through its membership in DataCite.

Examples
Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127–797. Geological Institute, University of Tokyo.
doi:10.1594/PANGAEA.726855
http://dx.doi.org/10.1594/PANGAEA.726855

SAFOD (2008): SAFOD Main Hole downhole logging data phase 2 (2005), 3387–3799m. Scientific Drilling Database.
doi:10.1594/GFZ.SDDB.1128
http://dx.doi.org/10.1594/GFZ.SDDB.1128

B. Kirchhof (2009) Silicone oil bubbles entrapped in the vitreous base during silicone oil removal, Video Journal of Vitreoretinal Surgery.
doi:10.3207/2959859860
http://dx.doi.org/10.3207/2959859860

Resources
DataCite
http://datacite.org

ICPSR Citation Recommendations
https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/citations.html

Purdue Online Writing Lab (OWL) Citation Resources
http://owl.english.purdue.edu/owl/section/2/

5 Archiving

5.1 Preservation issues and strategies

The active preservation of data is necessary to ensure its long-term availability and utility. Effective data archiving and preservation goes beyond data storage issues. It centers on developing strategies to safeguard the data and to ensure its fitness for contemporary purposes. Preservation strategies should include identifying risks to the data (obsolescence, corruption, deterioration, etc.) and how they will be addressed, defining what other content or information will need to be captured and associated with the data to ensure that it can be rendered and understood in the future, and consideration for how data will be submitted to, managed in, and retrieved from a preservation system.

Although preservation takes place at the end of the data's lifecycle, effective data preservation requires planning and preparation, which should begin at the project development stage.

Possible Questions to Consider

How will the preservation of your data serve the needs of likely user communities?

The decisions made and actions taken to preserve your research data should be driven by the needs of the communities that will make use of the data. These needs should be clearly identified to the extent possible and incorporated as central components of your preservation strategy and policy.

How will the data need to be prepared for preservation?

Properly preserving digital data requires that it will retrievable, understandable and usable in the future. This will often require additional information about the data to be generated or preserved alongside of the data. For example, information about the software used to produce or analyze the data, or lab notebooks or communications between project personnel could be needed to provide context for the data. You may need to provide documentation on missing data or any anomalies in your data set as well. Data that contain confidential or sensitive information may require additional preparation to ensure that information that should not be shared openly is protected from disclosure.

What resources will need to be acquired, or what arrangements will need to be made, in order to enable the preservation of the data?

If you are planning on preserving the data locally, you will need to consider what resources you will need and how you will obtain them. If you will be transferring your data to a 3rd party to preserve your data, you should develop plans for how and when you will transfer the data to this party. You should also consider any requirements or conditions made by the 3rd party in accepting your data. The roles and responsibilities of personnel in preserving the data should be identified and accounted for both during and after the lifespan of the research project.

For help in finding information relating to preserving and archiving research data, contact a Purdue Libraries' faculty in your subject area. A list of faculty librarians is available here.

Resources

The Digital Preservation Handbook - Produced by the Digital Preservation Coalition
http://www.dpconline.org/advice/preservationhandbook

Digital Preservation Management Tutorial
http://www.icpsr.umich.edu/dpm/index.html

Digital Preservation (Library of Congress) - Information and resources about digital preservation from the Library of Congress.
http://www.digitalpreservation.gov/

5.2 Selection and Appraisal for Preservation

Not all data needs to be preserved indefinitely, or even at all. Data that can easily be reproduced or are readily available from other sources may not require preservation for example. Researchers should consider which of the data they will generate will have lasting value to their research communities or other constituencies as a part of their preservation planning. Articulating criteria for selection and appraisal of research data is best done early on in the project, ideally in the planning stages.

> **Example** - ICPSR, a repository for social science data, uses the following criteria in determining whether or not to accept a data set into their collection:
>
> - They have substantive current value for research and instruction.
> - They have enduring value.
> - They are unique in some way.
> - They are useful for the development of emerging research and statistical techniques.

Possible Questions to Consider

For what purpose(s) will your data be preserved? How would you imagine that your data would be used by others over time?

Some possible responses are: for re-use in future research projects, to enable the verification of research findings, or for the historical record. Keep in mind to consider the different perspectives and needs of potential audiences. Data may have value beyond the immediate research interests of your own field of research.

How long are your data likely to retain their value?

Data may lose value over time for a variety of reasons, such as when new technologies are employed or new data sets are generated. Considerations of how long the research data are likely to retain value should be framed from the purposes you have identified in preserving them.

For help in finding information related to the selection and appraisal of research data for archiving and preservation, contact the Purdue Libraries' Research Data Group.

Resources

The Selection of Research Data: Guidelines for Appraising and Selecting Research Data - A report that provides set of general guidelines for apprising and selecting research data for preservation and other purposes.
http://repository.tudelft.nl/view/ir/uuid%3Adbab8a19-542a-4c4d-96b4-df8cc39333db/ -