USGS
science for a changing world

Data Management Training Modules:

Planning for Data Management

Creating Data Management Plans for your Research Project

U.S. Department of the Interior
U.S. Geological Survey

Version 1, October 2013

Narration: Introduction Slide 1

Welcome to the USGS Data Management Training Modules, a three part training series that will guide you in understanding and practicing good data management. Today we will discuss Planning for Data Management – Creating Data Management Plans for your Research Project.

Module Objectives

- Define & describe data management plans (DMPs).
- Explain the benefits of creating a DMP.
- Provide instructions on how to prepare a DMP.
- Describe components of a DMP.

Narration: Slide 2

Module Objectives

In this module, we will provide an overview of data management plans. First, we will define and describe Data Management Plans, or DMPs. We will then explain the benefits of creating a DMP. Finally, we will provide instructions on how to prepare a DMP, including covering key components common to most DMPs.

From Flickr by Barbies Land

What is a data management plan?

A formal document that:

outlines data management considerations before work begins…

…and how data will be handled during and after the completion of research.

USGS

Additional Resources

Glossary

Narration: Slide 3

What is a Data Management Plan?

First, what is a data management plan? A data management plan, or DMP, is a formal document that outlines data management considerations before work begins and how data will be handled during and after the completion of research.

A DMP is **not** a research plan, and the distinction between the two is worth noting.

- A DMP supports the data handling and development aspects of the formal research plan, but is only one aspect of a research project.
- A research plan describes the purpose of the project, details about the work to be done, and the project management that will support it. This plan may include data management, but not necessarily.

It is a common misconception that if project management is being done then data management is also occurring, but in reality these are two related but different activities.

DMP for Funders:

- **A plan submitted alongside grant applications.**
- **An outline of:**
  - business requirements.
  - data sources.
  - data acquisition methods.
  - Standards.
  - data flows.
  - Metadata.
  - sharing/access.
  - long-term storage.
- **Includes the “why.”**

From Flickr by 401(K) 2013

USGS

Additional Resources

Glossary

Narration: Slide 4

DMP for Funders:

Funders often require DMPs alongside grant applications. These are a specific type of DMP, often shorter and less detailed than traditional DMPs.

Funder DMPs often contain the components listed here, including how the applicant will acquire data, use applicable standards, ensure adequate documentation, provide protection from loss, and share and preserve the data underpinning their research. In cases where the funder has existing data management policy that the applicant is not adopting or deviating from, such decisions should be justified. For example, limits on data sharing can be an issue if the applicant has policy conflicting with the funding entity.

The “why” should be covered for each of the components listed. For example, not only should the DMP describe which metadata standard will be used, but why that particular standard was chosen.

Why Prepare a DMP?

- **Saves time.**
- **Data are an investment.**
- **Increases research efficiency:**
  - Ensures you and others will be able to understand and use data in future.
  - Prevents duplication of effort.
- **Satisfies funding agency requirements.**

From Flickr by natalinha

USGS

Additional Resources

Glossary

Narration: Slide 5

Why Prepare a DMP?

- First and foremost, it saves time. Plan your work and then work your plan: preparing a DMP saves time in the long run! It is easier to plan how data will be acquired and managed through the project lifecycle at the outset as opposed to trying to organize or retrofit data after they have been collected or acquired.
- Secondly, collection and preparation of data is an investment of both time and money during the research process. Development of a DMP helps to support and ensure your investment will meet the needs of your research.
- A high-quality DMP also increases research efficiency. Work spanning several months or years can result in the loss of important details or information about the data if they are not well documented. If data are well documented over the course of a project it helps to mitigate problems in understanding and using the data later on by both you and others.
- In addition to considering the benefits and incentives for creating DMPs, consider that many funding agencies now require they be submitted.

Where Should you Begin?

- Document what you know now:
  - Keep it simple and understandable.
  - Use a template or metadata standard for identifying key information types.
- Data flow diagrams are useful and can help.
- Pinpoint potential issues early.
- Share the plan with your team.
- Avoid procrastination and immobilization.

From Flickr by jakeandlindsay

USGS

Additional Resources

Glossary

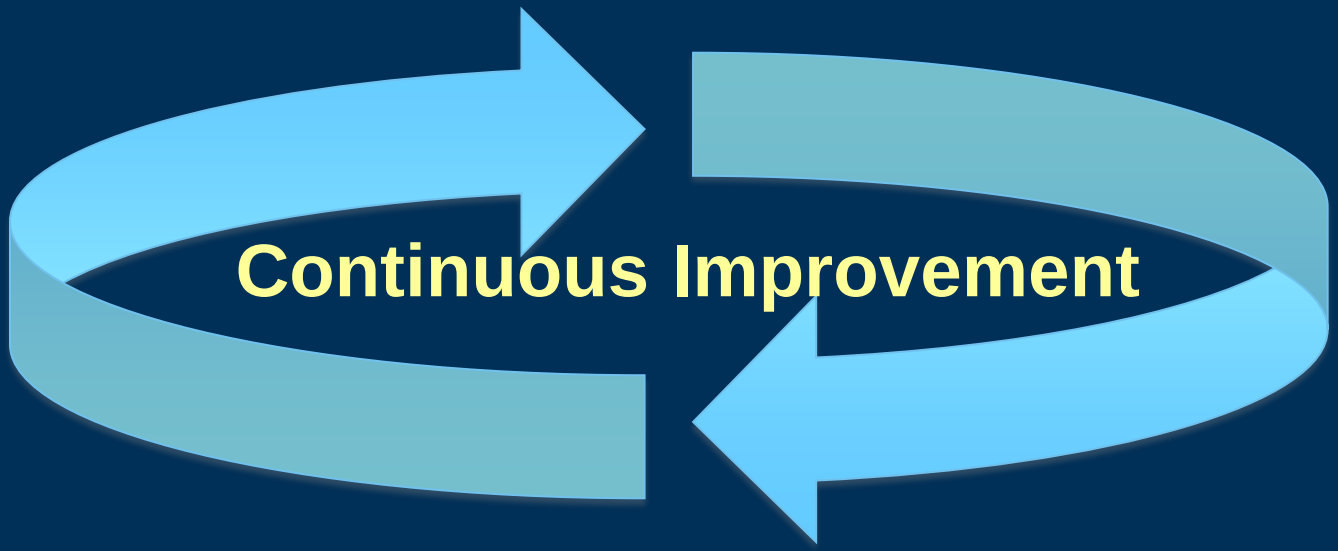
Narration: Slide 6

Where Should you Begin?

Documentation often improves and becomes more complete if you start start early and continue to add and improve over the course of a project. Keeping the language simple, for example, by defining acronyms and avoiding jargon decreases the likelihood of someone misunderstanding the information. Templates or existing documentation standards, such as metadata standards, can be useful as a way to standardize what is being documented for consistency and completeness. Tools such as data flow diagrams when created early on can be useful for visualizing the process through which new or existing data are acquired, prepared, and used in research. Make sure the research team is aware of the DMP and that everyone has an understanding of the expectations stemming from it.

Revisit and Continuously Improve

- Review your DMP regularly to ensure its contents are current.
- Use the plan as a guide for daily activities.



USGS

Additional Resources

Glossary

Narration: Slide 7

Revisit and Continuously Improve

The DMP should be thought of as a “living document.” That is, you should continuously add to, improve upon, and revisit it to make sure it’s up to date.

Research can be a trial and error process and as such the data management issues germane to the work may shift over time. Refer to the DMP often and use it as a guide for daily activities related to data management.

Components of a DMP

1. Information about data & data format.
2. Metadata content and format.
3. Policies for access, sharing and re-use.
4. Long-term storage and data management.
5. Budget.

USGS

Additional Resources

Glossary

Narration: Slide 8

Components of a DMP

A good DMP should address the 5 high-level areas of data management shown here. Depending on the research scope of work, some of these components may need more emphasis than others, but regardless the DMP should touch on these five major topics. The slides that follow will go into more detail for each.

1. Information About Data & Format

- Data types: newly collected and existing datasets.
- Business rules (temporal, spatial, other requirements).
- Conceptual data model.
- Data acquisition methods.
- Data and process flow model.
- Quality assurance and quality control measures.

USGS

biology.kenyon.edu

From Flickr by Lazurite

USGS

Additional Resources

Glossary

1. Information About Data & Format

First and foremost, the DMP should address issues related to data and data format. The types of data needed for the research should be outlined with a defined minimum amount of information for each dataset (for example, dataset name, description, applicable protocols, required software to access or use the data).

In a data management context business rules are typically expressed as constraints. These might include spatial scope and scale constraints such as a scale limit on geographic data or study area boundary, temporal scope and scale, or any other aspects that might influence design of the data or integrating one or more datasets.

A conceptual data model can be a useful tool to visualize the scope of the research by identifying the principal datasets of interest and any associations among them. Graphical software or a whiteboard are often useful for illustrating the conceptual data model for a project.

Information should be provided about how the data will be acquired. The DMP should specify the acquisition considerations for each dataset which may include dataset-specific protocols, chain of custody requirements, specifications for field or lab data collection forms, and a data dictionary defining data fields or attributes.

A data and process flow model can be useful to visualize the steps, processes, and connections among them necessary for collecting, preparing, transforming, and integrating data over the course of a project. This model should represent a diagram for how datasets will move through a project from their respective source to the final data products used for research. Ideally, the diagram will consider the various states of data in a project which may include raw data, reviewed and approved for use base data, derived data (data created from combining or transforming base data or other derived data), and the project database which serves as the integrated data that will be used to conduct the research.

Quality assurance and quality control measures should also be addressed. This information should describe what steps will be taken to ensure that the datasets are free of error, issues, or mistakes. Quality assurance focuses on the processes used to create the data and the proactive steps used to ensure quality. The purpose of quality control is to identify errors in, for example, in a finalized dataset.

1. Information About Data & Format (cont.)

- Standard data and research terms and definitions.
- Data management roles and responsibilities.
- Security: version control, backup procedures, access restrictions.
- Data management work plan.
- Deliverable data products.

1. Information About Data & Format (cont.)

Your DMP should serve as a centralized reference by describing the nomenclature, terms, and definitions associated with the data. This includes acronyms, abbreviations, and technical terms both within the research or those that are applicable but have a broader scope that the research should adopt. Defining terms (for example, what does the term database mean in the context of the research DMP) helps to make clear the focus of the language in the DMP and helps to avoid confusion among those conducting the research.

The DMP should also outline the data management roles and the person(s) responsible for them. Individuals often “wear the hat” of one or more of the roles outlined in a research DMP. Assigning functions by role is useful to help focus as well as set boundaries among those involved with the project. Much like the conceptual data model, integrating in the team component by tying members of the research team to the defined roles helps to ensure that key functions will be performed. This also helps to create accountability and set expectations so that the assumption is not made that “someone else is handling that issue”. Common roles associated with data management can include: principle investigator, subject matter expert, data steward, data manager, data analyst, metadata specialist, data integration specialist, database administrator, GIS specialist, and application developer.

Security and protection of data assets should also be considered in a DMP. This can include where data and project files are stored; file access, encryption, and backup strategies; and data versioning procedures. Versioning procedures are especially important to consider with research teams consisting of multiple people accessing the same files or areas of a data architecture.

A data management work plan is useful for outlining the activities need to implement the associated DMP. The data management work plan is not a research work plan but should be developed in conjunction with a research work plan to ensure that mutual activities are complementary both with their timelines and how one supports the other. The activities in a data management work plan can vary greatly but may include creating data models, coordinating acquisition of existing data and signing of usage affidavits, creating programs and scripts to prepare data, the scheduling of field equipment necessary for data collection, the archival of datasets, the establishment of processes for backing up data and the review of any automated backups to ensure they completed correctly, and the installation and configuration of data management hardware and software.

Finally, a list of the data products that are expected to result from the project should be described.

2. Metadata Content & Format

Metadata is the documentation describing all aspects of the data (e.g., who, why, what, when, where).

- **What information should be included?**
  - **Any details that make data understandable and usable.**
- **How metadata will be created and/or captured.**
  - **Lab notebooks? GPS units? Web forms? Auto-saved on instrument? Manually entered?**
- **What format will be used for the metadata?**
  - **Standards for community (EML, ISO 19115, etc.).**
  - **Justification for format chosen.**

USGS

Narration: Slide 11

2. Metadata Content & Format

Metadata provides important information about a dataset. This includes who created the data, the content of the data, and the when, where, how, and why the data will be collected or created. Metadata records and captures critical information about a dataset and should be as thorough as possible. Describe what metadata will be needed to make sure the dataset is usable. How will these metadata be captured or created? This might be done via lab notebooks, instruments, web forms, or other means. Finally, describe the metadata standard that will be used and why it was chosen.

3. Policies for Access, Sharing, Reuse

- Obligations for sharing?
- Details of data sharing: When? How access can be gained?
- Ethical/privacy issues with data sharing.
- Intellectual property & copyright issues:
  - Institutional policies.
  - Funding agency policies.
  - Political or commercial license considerations.
- Citation:
  - How should data be cited when used?
  - Persistent citation?

Narration: Slide 12

3. Policies for Access, Sharing, Reuse

The third major component of a DMP is policies and provisions for data access, sharing, and reuse.

First, consider whether there are any existing obligations for sharing. Some funders or institutions require that datasets be shared, and this should be mentioned in the DMP.

You should include any important details about how the data will be shared, such as when will the data be made available, or how can others gain access.

There may be ethical or privacy issues related to sharing datasets, for example if endangered species or protected areas are involved, or if there are human subjects. These issues should be addressed in the DMP.

Intellectual property and copyright issues will depend on institutional and funder policies. These should be considered and included in your plans for access and sharing.

And finally, include how others should cite your data when they use it.

4. Long-Term Storage & Management

- What data will be preserved for the long term? For how long?
- Where will data be preserved?
- What metadata will be submitted alongside the datasets?
- Who will be responsible for preparing data for preservation? Who will be the main contact person for the archived data?

From Flickr by theManWhoSurfedTooMuch

USGS

Additional Resources

Glossary

Narration: Slide 13

4. Long-Term Storage & Management

When planning for data management, it’s wise to establish plans for long-term storage early and let the plan guide management efforts.

Not all data need to be kept: consider what data being collected or created will be preserved, and for how long. Establish what repository or data center will house the data, and whether there are any metadata standards or other requirements that should be considered during data collection. Determine what personnel will be responsible for both data preservation and for maintaining contact with the data center in the long term.

5. Budget

- Anticipated costs:
  - Data acquisition costs and fees
  - Time for data preparation & documentation.
  - Hardware and software for data preparation & documentation.
  - Personnel.
  - Archiving.
- How costs will be paid.

USGS

Additional Resources

Glossary

Narration: Slide 14

5. Budget

Finally, any DMP should include details on anticipated costs associated with data management and how those costs will be covered. These costs should be included in the budget of any research proposal, since data management is integral to the success of any research project. Costs that should be included are data acquisition, personnel time for data preparation and documentation, hardware and software necessary to manage the data, and archiving costs.

Key Points

- DMPs describe plans for managing data throughout the research project.
- There are many benefits to creating and maintaining DMPs.
- DMPs should be revisited frequently and adjusted as needed.
- Main DMP components are information about data and data format, metadata, data policies, long-term storage plans, and budgeting.

Key Points

In summary, a data management plan, or DMP, is a formal document that outlines data management considerations before work begins and how data will be handled during and after the completion of research. DMPs describe plans for managing data throughout the research project. A DMP provides many benefits including it helps save time in the long run, it helps you preserve your investment of time and money, it helps increase research efficiency, and it helps satisfy the requirement that many funding agencies now have for proposed research. Revisiting DMPs throughout the research process helps to ensure that it is up-to-date and also helps to keep daily activities focused. Finally, covering at a minimum the five main areas common to nearly all DMPs helps to ensure that the data associated with your research have been adequately planned for.