Data Management Planning

Failing to plan is planning to fail.

Alan Lakein

Data Management Planning Topics

1. What is a data management plan (DMP)?
2. Why prepare a DMP?
3. Components of a DMP
4. Example of a NSF DMP
5. Exercise

What is a DMP?

- A document that describes what you will do with your data **during** and **after** you complete your research

Why Prepare a DMP?

- Saves time

- 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

NSF DMP Requirements

From Grant Proposal Guidelines:

Plans for data management and sharing of the products of research. **Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan".** This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (in AAG), and may include:

1. the **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project

2. the **standards to be used for data and metadata** format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)

3. **policies for access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements

4. policies and **provisions for re-use**, re-distribution, and the production of derivatives

5. **plans for archiving** data, samples, and other research products, and for preservation of access to them

Components of a DMP

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

1. Information About Data & Data Format

1.1 Description of data to be produced

Experimental, Observational, Raw or derived,

Physical collections, Models, Images, etc.

1.2 How data will be acquired

When? Where? Methods?

1.3 How data will be processed

Software used, Algorithms, Workflows

1.4 File formats

csv, tab-delimited, naming conventions

1.5 Quality assurance & quality control

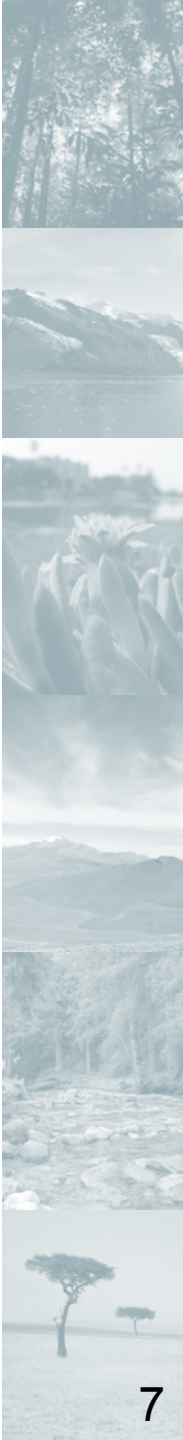1.6 Existing data

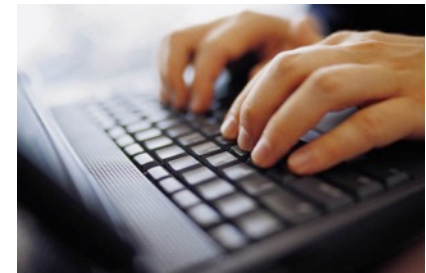*If existing data are used, what are its origins?*

Will your data be combined with existing data?

What is the relationship between your data and existing data?

1.7 How data will be managed in short-term

Version control, Backing up, Security & protection,

Who will be responsible?

2. Metadata Content & Format

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

2.1 What metadata are needed

   Any details that make data understandable and usable

2.2 How metadata will be created and/or captured

   Lab notebooks? GPS units?

   Auto-saved on instrument? Manually entered?

2.3 What format will be used for the metadata

   Standards for community (EML, ISO 19115, etc.)

   Justification for format chosen

3. Policies for Access, Sharing, Reuse

3.1 Obligations for sharing

- Funding agency, institution

3.2 Details of data sharing

- How long?
- When?
- How access can be gained?

3.3 Ethical/privacy issues with data sharing

3.4 Intellectual property & copyright issues

- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.5 Intended future uses/users for data

3.6 Citation

- How should data be cited when used?
- Persistent citation?

4. Long-term Storage & Data Management

4.1 What data will be preserved

4.2 Where will it be preserved

Most appropriate archive for data

Community standards

4.3 Data transformations/formats needed

Consider archive policies

4.4 Who will be responsible

Contact person for archive

5. Budget

5.1 Anticipated costs

Time for data preparation & documentation

Hardware/software for data preparation & documentation

Personnel

Archive costs

5.2 How costs will be paid

Data in Real Life: A DMP Example (1)

Project name: Effects of temperature and salinity on population growth of the estuarine copepod, *Eurytemora affinis*

Project participants and affiliations:

> Carly Strasser (University of Alberta and Dalhousie University)
>
> Mark Lewis (University of Alberta)
>
> Claudio DiBacco (Dalhousie University and Bedford Institute of Oceanography)

Funding agency: CAISN (Canadian Aquatic Invasive Species Network)

Description of project aims and purpose:

> We will rear populations of *E. affinis* in the laboratory at three temperatures and three salinities (9 treatments total). We will document the population from hatching to death, noting the proportion of individuals in each stage over time. The data collected will be used to parameterize population models of *E. affinis*. We will build a model of population growth as a function of temperature and salinity. This will be useful for studies of invasive copepod populations in the Northeast Pacific.

Video Source: Plankton Copepods. Video. E*ncyclopædia Britannica Online*. Web. 13 Jun. 2011

Data in Real Life: A DMP Example (2)

1. Information about data

Every two days, we will subsample *E. affinis* populations growing at our treatment conditions. We will use a microscope to identify the stage and sex of the subsampled individuals. We will document the information first in a laboratory notebook, then copy the data into an Excel spreadsheet. For quality control, values will be entered separately by two different people to ensure accuracy. The Excel spreadsheet will be saved as a comma-separated value (.csv) file daily and backed up to a server. After all data are collected, the Excel spreadsheet will be saved as a .csv file and imported into the program R for statistical analysis. Strasser will be responsible for all data management during and after data collection.

Our short-term data storage plan, which will be used during the experiment, will be to save copies of 1) the .txt metadata file and 2) the Excel spreadsheet as .csv files to an external drive, and to take the external drive off site nightly. We will use the Subversion version control system to update our data and metadata files daily on the University of Alberta Mathematics Department server. We will also have the laboratory notebook as a hard copy backup.

Data in Real Life: A DMP Example (3)

2. **Metadata format & content**

    We will first document our metadata by taking careful notes in the laboratory notebook that refer to specific data files and describe all columns, units, abbreviations, and missing value identifiers. These notes will be transcribed into a .txt document that will be stored with the data file. After all of the data are collected, we will then use EML (Ecological Metadata Language) to digitize our metadata. EML is on of the accepted formats used in Ecology, and works well for the type of data we will be producing. We will create these metadata using Morpho software, available through the Knowledge Network for Biocomplexity (KNB). The documentation and metadata will describe the data files and the context of the measurements.

Data in Real Life: A DMP Example (4)

3. Policies for access, sharing & reuse

We are required to share our data with the CAISN network after all data have been collected and metadata have been generated. This should be no more than 6 months after the experiments are completed. In order to gain access to CAISN data, interested parties must contact the CAISN data manager (data@caisn.ca) or the authors and explain their intended use. Data requests will be approved by the authors after review of the proposed use.

The authors will retain rights to the data until the resulting publication is produced, within two years of data production. After publication (or after two years, whichever is first), the authors will open data to public use. After publication, we will submit our data to the KNB allowing discovery and use by the wider scientific community. Interested parties will be able to download the data directly from KNB without contacting the authors, but will still be encouraged to give credit to the authors for the data used by citing a KNB accession number either in the publication's text or in the references list.

Data in Real Life: A DMP Example (5)

4. **Long-term storage and data management**

    The data set will be submitted to KNB for long-term preservation and storage. The authors will submit metadata in EML format along with the data to facilitate its reuse. Strasser will be responsible for updating metadata and data author contact information in the KNB.

5. **Budget**

    A tablet computer will be used for data collection in the field, which will cost approximately $500. Data documentation and preparation for reuse and storage will require approximately one month of salary for one technician. The technician will be responsible for data entry, quality control and assurance, and metadata generation. These costs are included in the budget in lines 12-16.

Exercise:

http://dmp.cdlib.org/

What can the DMP Tool do for you?

What a researcher can do:

View sample plans

Preview funder requirements

Create, save, edit, publish plan

View, use, past plans

Use help (generic and institution specific)

View news and latest changes

Data Management Planning (Part 2)

A goal without a plan is just a wish.

Larry Elder

Data Management Planning Tool

DMPTool
Guidance and Resources for your Data Management Plan

NSF-GEN: Generic: Plan description

The NSF-GEN: Generic plan will cover the subject areas listed to the left.

You can save a plan in progress and return later to finish or edit.

Progress

Click on a section below to edit it at any time.

⊘ = complete

Plan description

1. Types of data produced
2. Data and metadata standards
3. Policies for access and sharing
4. Policies for re-use, redistribution
5. Plans for archiving & preservation

Plan Name: (required)
Please give your plan a name to help you identify it in the future

Solicitation Number:

Comment:
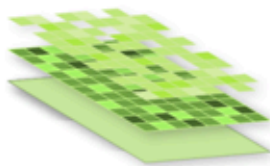Provide any notes you want to appear on your My Plans page. This will not appear in the document

B *I* U abe X₂ X² | [icons] | [icons] | [icons] | [icons]

Cancel   Save and next

Resources

General
NSF Data Sharing Policy

NSF Data Management Plan Requirements

DMP Tool
Build your data management plan

Home   About DMP Tool ▾   DMP News   My Plans   Help ▾

Progress

Sections marked with a check are complete. You can navigate to a section and edit at any time.

NSF-GEN: Generic

Cover page

⊘ 1. Types of data produced
⊘ 2. Data and metadata standards
⊘ 3. Policies for access and sharing
⊘ 4. Policies for re-use, redistribution
⊘ 5. **Plans for archiving & preservation**

Plans for archiving data, samples, and other research products, and for preservation of access to them.

> **Suggested answer text; copy and paste as needed:**
>
> As advised by University of Virginia Library staff members, I plan on depositing my research data in the UVA institutional repository – Libra. I will submit the necessary metadata and other resources to make my data accessible for future users. In accordance with the University of Virginia policy RES-002, "Policy: Laboratory Notebook and Recordkeeping," the data will be preserved for a minimum of five years upon completion of the project. However the current preservation plan for Libra will be to preserve the data indefinitely. The Libra backup plan provides for data redundancy including off-site storage.

As advised by University of Virginia Library staff members, I plan on depositing my research data in the UVA institutional repository – Libra. I will submit the necessary metadata and other resources to make my data accessible for future users. In accordance with the University of Virginia policy RES–002, "Policy: Laboratory Notebook and Recordkeeping," the data will be preserved for a minimum of five years upon completion of the project. However the current preservation plan for Libra will be to preserve the data indefinitely. The Libra backup plan provides for data redundancy including off–site storage.

Resources

University of Virginia
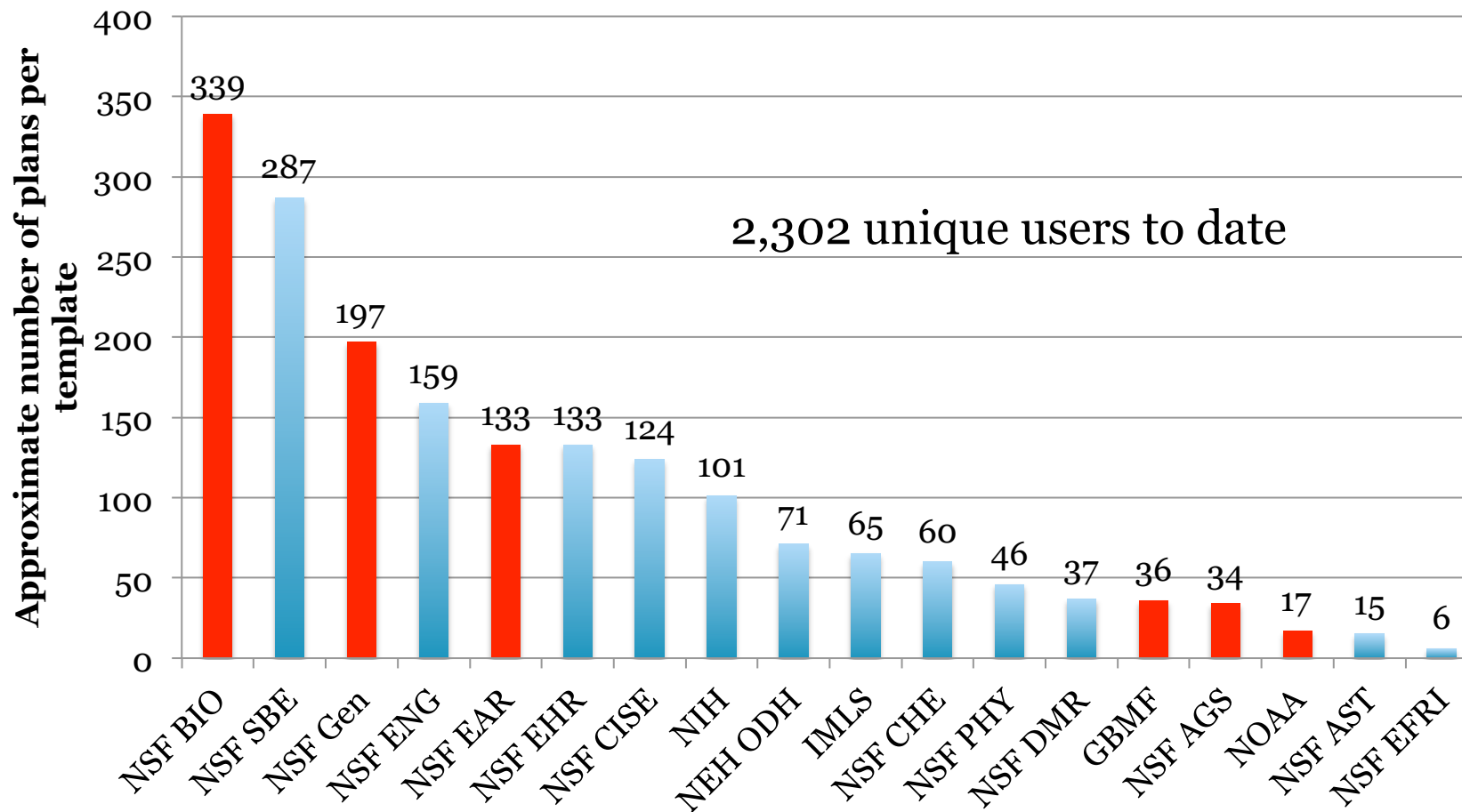UVa Scientific Data Consulting Group

Archiving & Sharing Data Guidance

UVa Policy RES-002: "Laboratory Notebook and Recordkeeping"

General
NSF Data Sharing Policy

NSF Data Management Plan Requirements

Plans per template (as of June 2012)

2,302 unique users to date

http://dmptool.org

Build your Data Management Plan

Data Management Planning Topics

1. When do you plan?
2. How do you develop a plan?
3. Who does the planning?

When do you plan?

At the beginning … and throughout the research life cycle

From: www.jisc.ac.uk

How do you plan?

- View DM planning as a part of the research process

- "The better is the enemy of the 'good enough'"

  - Don't hesitate because you don't know all the answers

    - Use metadata from your proposal, publications, etc.

- Review and revise routinely

Who plans?

DMP Resources

http://dmp.cdlib.org/

http://dataone.org

Thanks to Carly Strasser, Amber Budden, Viv Hutchison, and DMP Tool contributors!