Developing Data Management Plans

Mozilla Open Science Workshops

(Uganda, Ethiopia, Sudan, and Kenya)

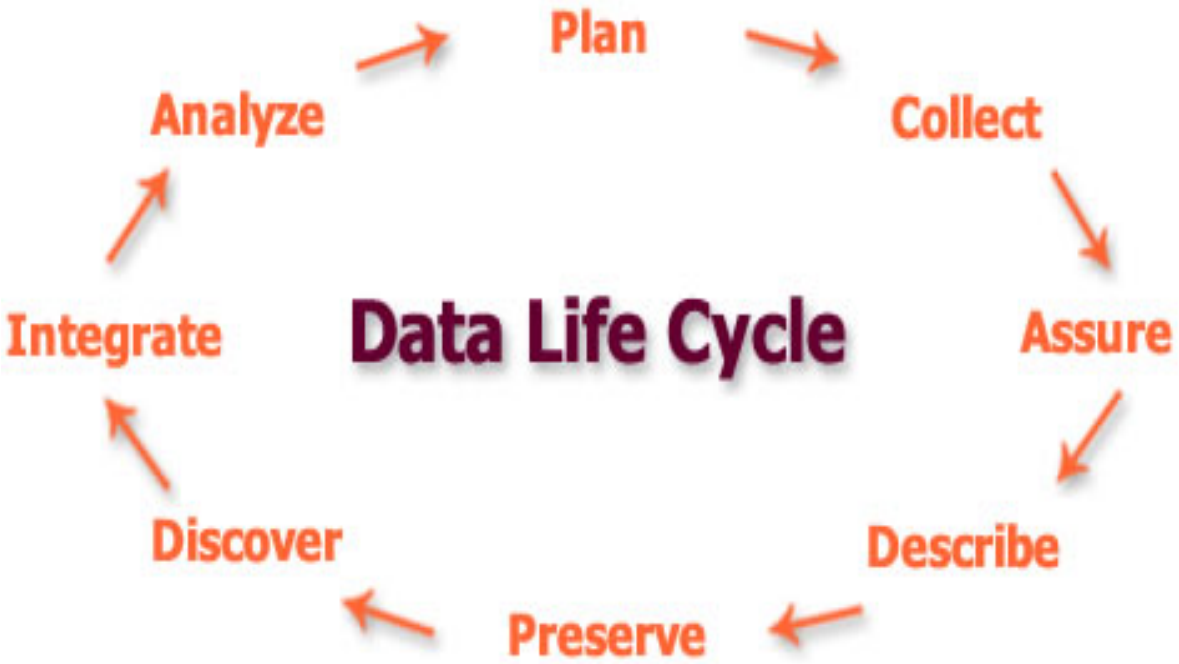
12-22 November 2018

Dr. Amel Ghouila, Prof. Faisal Mohamed Fadlelmola

1. H3ABioNet, Institut Pasteur de Tunis, Tunisia

2. H3ABioNet, Centre for Bioinformatics & Systems Biology, University of Khartoum, Sudan

Data life cycle



Source: DataONE's "Best Practices"

What is Data Management Plan?

- A document that tells how a researcher will
  - collect,
  - document,
  - describe
  - share
  - preserve

data that will be generated as part of a project.

Why Manage Your Research Data?

Managing your research data will help you:

- Ensure long-term preservation of data;
- Encourage the discovery and use of your data to explore new research questions;
- Encourage the discovery and use of your data to explore new research questions;
- Improve your data's accuracy, completeness, and usability;
- Align/Comply with ethics and privacy policies.
- Meet funding agency requirements;
- Write more competitive grant applications;

Components of a Typical DMP

1. Types of data to be collected or produced and the processes used
2. Data formats and metadata
3. Access, sharing and privacy :
   - How to access the data,
   - Information about privacy and/or intellectual property;
4. Policies and guidelines for data re-use and re-distribution,
5. Data storage and preservation
   - How to ensure long term preservation and access,
   - Who is going to be responsible for managing the data for the project duration;

Metadata

- Metadata is '*data about data*'.
- Information necessary to make your data '**independently understandable**'.
- Information including how was the data created, analyzed and stored
- Using established metadata standards will help make your data **discoverable**, **citable**, and **ready-to-use** by others.

(National Information Standards Organization, 2004).

Basic Metadata Elements

Title; Creator; Date Created; Format; Subject; Unique Identifier (*ideally, a Digital Object Identifier, or DOI*); Description of the specific data resource; Coverage (*spatial or temporal*); Publishing Organization; Type of Resource; Rights/Licensing/Ethics approval; Funding/Granting Agency

Many metadata Standards (specific to repositories or fields)

Biomedical Metadata

- *Reagent Metadata*: Information about the clinical samples, biological reagents (e.g. cell lines, antibodies, siRNAs), chemical reagents (e.g. drugs), etc. used to generate the data.
- *Technical Metadata*: Information automatically generated by research instruments and associated software.
- *Experimental Metadata*: Information about the experimental conditions (e.g. assay type, time points), the experimental protocol, and the equipment used to generate the data.
- *Analytical Metadata:* Information about data analysis methods including software name and version, quality control parameters, and output file type details.
- *Dataset Level Metadata*: Information about the objectives of the research project, participating investigators, relevant publications, and funding sources.

Data Repositories & Archives

- Many discipline-specific repositories
- Re3data.org Registry of Research Data Repositories
- The European Genome-phenome Archive (EGA): permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects
- The Simmons Open Access Directory (OAD) provides a growing list of open access data repositories

Best Practices in Developing a DMP

- Writing a data management plan is increasingly seen as a key part of the academic research process.

- The following practices are fundamental to effective data management and can be applied to all disciplines:
  - ❑ The Data Management Plan;
  - ❑ Data Storage;
  - ❑ Data Documentation;
  - ❑ Ethical Issues;
  - ❑ Sharing Data.

The Data Management Plan

- Adhere to the guidelines set by any funding agencies and institutions that are sponsoring the research.
- Complete your DMP early so that it will not be sidelined when you start collecting data.
- The minimum expenses to include when calculating your data management costs are: data creation, processing, analysis, storage, sharing, and preservation.
- Remember that some Funding Agencies accept these costs in grant applications --- be sure to include these costs.

Data Storage

- To prevent data from being lost due to incompatibility, store it **in formats** and **on hardware** that are open standard

- Be **consistent and descriptive** when naming files in your data set
- **Backup your data** on local and remote external servers and have a contingency plan for restoring lost data.

Data Documentations

- Use metadata to record details about a study such as:
  - its context
  - the dates of data collections
  - data collection methods, etc.

- Use an established metadata standard appropriate to your field
- Examples:
  - For Social Science data: DDI is the standard most often used.
  - For Ecological data: Ecological Metadata Language (EML)
  - Geographic Data: The Federal Geographic Data Committee has a metadata schema for digital geospatial data.

Ethical Issues

- All sensitive information in your data should be redacted (removed) before depositing in a public archive or repository.
- Access to data may need to be embargoed (limited for a certain amount of time) in order to ensure privacy.
- Be aware of the ownership and intellectual property rights concerning your data.
- Researchers should be reminded to seek 'up front' consent from study participants to archive data collected about them:
- Check institution/country policies

Ethical Issues: some resources

- Health Sciences Research Ethics Board (HSREB)

- Human-subject data collection:

Recommended Informed Consent Language for Data Sharing

Sharing Data

- Sharing data is an essential part of the Data Life Cycle.
- Data can be shared informally with other researchers or posting it to a website.
- Informal methods of sharing data make it difficult to find and access it in both the short and the long-term.
- Depositing your data in an appropriate data archive helps ensuring preservation and re-use of your data.

DMP: Other Considerations

What about finishing my research?

- You can set an embargo date on deposited data to prevent others from having access until after your research is complete and published.
- In the meantime, your data is safe, well-documented, and available exclusively to you and your research team.

Do I retain Copyright?

- Depositing data into a repository does not generally affect copyright ownership.
- Depositors can specify conditions that secondary users must adhere to when accessing deposited data.
- Data are normally shared for research and teaching purposes only -- not for commercial purposes.

What kinds of research data can be deposited?

- A wide variety of data types and sources from all disciplines.

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Data description	A description of the information to be gathered; the nature and scale of the data that will be generated or collected.	Yes	Expected Data
Existing data	A survey of existing data relevant to the project and a discussion of whether and how these data will be integrated.	Yes	Expected Data
Format	Formats in which the data will be generated, maintained, and made available, including a justification for the procedural and archival appropriateness of those formats.	Yes	Data Format and Dissemination
Metadata	A description of the metadata to be provided along with the generated data, and a discussion of the metadata standards used.	Yes	Data Format and Dissemination

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Storage and backup	Storage methods and backup procedures for the data, including the physical and cyber resources and facilities that will be used for the effective preservation and storage of the research data.	Yes	Data Storage and Preservation of Access
Security	A description of technical and procedural protections for information, including confidential information, and how permissions, restrictions, and embargoes will be enforced.	Yes	Data Format and Dissemination
Responsibility	Names of the individuals responsible for data management in the research project.	Yes	Roles and Responsibility

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Intellectual property rights	Entities or persons who will hold the intellectual property rights to the data, and how IP will be protected if necessary. Any copyright constraints (e.g., copyrighted data collection instruments) should be noted.	Yes	Data Format and Dissemination
Access and sharing	A description of how data will be shared, including access procedures, embargo periods, 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.	Yes	Data Storage and Preservation of Access

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Audience	The potential secondary users of the data.	Yes	Data Format and Dissemination
Selection and retention periods	A description of how data will be selected for archiving, how long the data will be held, and plans for eventual transition or termination of the data collection in the future.	Yes	Period of Data Retention
Archiving and preservation	The procedures in place or envisioned for long-term archiving and preservation of the data, including succession plans for the data should the expected archiving entity go out of existence.	Yes	Data Storage and Preservation of Access

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Ethics and privacy	A discussion of how informed consent will be handled and how privacy will be protected, including any exceptional arrangements that might be needed to protect participant confidentiality, and other ethical issues that may arise.	Yes	Data Format and Dissemination
Budget	The costs of preparing data and documentation for archiving and how these costs will be paid. Requests for funding may be included.		
Data organization	How the data will be managed during the project, with information about version		

Elements of a Data Management Plan

Element	Description	Recommended?	Main Sections
Quality Assurance	Procedures for ensuring data quality during the project.		
Legal requirements	A listing of all relevant federal or funder requirements for data management and data sharing.		

Writing a DMP Online Tools

- The **Portage Network** has developed an online tool designed to help Canadian academic researchers develop and implement research data management plans (http://portagenetwork.ca/).

- The **DMP Assistant** provides a step-by-step approach to writing a data management plan, with step-by-step guidance provided along the way (https://assistant.portagenetwork.ca/en).

- The **DMPTool**, includes the steps necessary to create data management plan for one of several funders, including the NSF (https://dmptool.org/).

The DMPTool

- Creates ready-to-use data management plans for specific funding agencies;
- Meets funder requirements for data management plans;
- Gets step-by-step instructions and guidance for your data management plan as you build it;
- Learn about resources and services available at your institution to help fulfill the data management requirements of your grant.

Ten Simple Rules for Creating a DMP

- Determine the Research Sponsor Requirements
- Identify the Data to Be Collected
- Define How the Data Will Be Organized
- Explain How the Data Will Be Documented
- Describe How Data Quality Will Be Assured
- Present a Sound Data Storage and Preservation Strategy
- Define the Project's Data Policies
- Describe How the Data Will Be Disseminated
- Assign Roles and Responsibilities
- Prepare a Realistic Budget

Source: Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11(10): e1004525. https://doi.org/10.1371/journal.pcbi.1004525

Exercise: Developing a DMP

1st Project for developing a DMP: Biology Background

Title: Atmospheric CO2 Concentrations, Mauna Loa Observatory, Hawaii, 2011--2013.

Purpose of project: The purpose of this proposed project is to study the controls on the concentration of atmospheric CO2 using high precision and accuracy measurements at a remote island observatory. We propose to measure the concentrations of CO2 in the atmosphere at the Mauna Loa Observatory, Hawaii. The methodology for sample collection and analysis during this project will generate highly accurate and precise data that can be seamlessly added to the existing Mauna Loa CO2 record (1958-2010) [1,2]. A major theme for this project is to identify and minimize systematic measurement errors through rigorous sampling and calibration procedures. We have chosen to use an iconic data product — the Keeling Mauna Loa CO2 record—for this example Data Management Plan. All environmental scientists are familiar with this data record. It is posted in the atrium of the U.S. National Academy of Sciences, next to the DNA Double helix. We are writing this Data Management Plan as if it were to be included in an NSF proposal in 2011.

2nd Project for developing a DMP: Computing Background

Title: Improving the long-term preservability of HDF--formatted data by creating maps to file contents.

Purpose of project: This proposed project will create ancillary metadata that will enable future users to read all HDF (Hierarchical Data Format) formatted data. Currently, users must have an HDF software library to read the data, or Employ a tool that uses that library. It is uncertain how long the library, which is required to read the files, will be in existence. Loss of this library would mean the loss of all data in HDF format. Currently there are about 800 different types of HDF--formatted files for remote sensing data at NASA alone. This stewardship project will protect against loss of the HDF software libraries

currently required to read the HDF files.

The HDF Group (www.hdfgroup.org), a non---profit company whose mission is to sustain the HDF technologies and support HDF user communities worldwide, is developing the map generation software and the specification or schema for maps for HDF data files.

Conclusions

For writing an effective Data Management Plan, provide a general description of the data expected to be produced over the course of the project. Consider the following:

- A brief, non-technical description of the data (including code or software, if appropriate) the project will produce. This should be a non-technical description that provides a very general idea of what data will be generated throughout the research project.
- If the proposed research involves obtaining data from other sources, provide a brief description of the data including its content, source, and any particular conditions for obtaining and using the data. Describe plans for redistributing any derived data products, if applicable.
- Indicate which data you will share and at what stage (raw, processed, reduced, or analyzed).
- Describe why the data you will share will be of interest to a broader community and how your plan will maximize the potential for reuse of the data.

Conclusions

- A DMP provides an easy-to-follow road map that will guide and explain how data are handled throughout the project and after the project is completed.
- Useful for collaborators and funders

Acknowledgements

- Dr. Sumir Panji, co-lead of the H3ABioNet DMP project.
- Mozilla Science Group for funding the “Africa Wow Tour” workshops in East Africa.
- H3ABioNet
- Local workshop organizers.