DataONE

www.dataone.org

Primer on Data Management: What you always wanted to know*

* but were afraid to ask

Carly Strasser, Robert Cook, William Michener, Amber Budden

Contents

1. Objective of This Primer

The goal of data management is to produce self-describing data sets. If you give your data to a scientist or colleague who has not been involved with your project, will they be able to make sense of it? Will they be able to use it effectively and properly? This primer describes a few fundamental data management practices that will enable you to develop a data management plan, as well as how to effectively create, organize, manage, describe, preserve and share data.

2. Why Manage Data?

2.1. It will benefit you and your collaborators

Establishing how you will collect, document, organize, manage, and preserve your data at the beginning of your research project has many benefits. You will spend less time on data management and more time on research by investing the time and energy before the first piece of data is collected. Your data also will be easier for you to find, use, and analyze, and it will be easier for your collaborators to understand and use your data. In the long term, following good data management practices means that scientists not involved with the project can find, understand, and use the data in the future. By documenting your data and recommending appropriate ways to cite your data, you can be sure to get credit for your data products and their use [1].

2.2. It will benefit the scientific community

Although data management plans may differ in format and content, several basic elements are central to managing data effectively. Ideally, data should be managed so that any scientist (including the collector or data originator) can discover, use and interpret the data after a period of time has passed. A key component of data management is the comprehensive description of the data and contextual information that future researchers need to understand and use the data. This description is particularly important because the natural tendency is for the information content of a data set or database to undergo entropy over time (i.e. data entropy), ultimately becoming meaningless to scientists and others [2].

An effective data management program would enable a user 20 years or longer in the future to discover, access, understand, and use particular data [3]. This primer summarizes the elements of a data management program that would satisfy this 20-year rule and are necessary to prevent data entropy.

2.3. Journals and sponsors want you to share your data

Research data are valuable products of the scientific enterprise that historically have not been well preserved or archived. Sponsors [4] and scientific journals [5] are now encouraging or requiring sound data management and data sharing. Effective data management practices are therefore critical to the scientific process.

Government agencies are under increasing pressure to demonstrate the benefits of the research they sponsor, both in terms of scientific findings (published papers) as well as data products. For instance, a 2007 US Government and Accounting Office Report summarized the issues associated with the loss of individual investigators' data and how this data loss deprives science and society of many of the benefits of research [6].

In January 2011, the National Science Foundation (NSF) instituted the requirement that a two-page data management plan be included in every grant proposal as a supplemental document [7]. Some individual NSF Directorates, Divisions, and Programs provide more specific guidelines, however NSF is generally relying on scientists from the various disciplines it supports to set expectations for data management through the peer-review process.

3. How To Use This Primer

This document highlights the basics of data management. It provides guidance about how to organize, manage, and preserve your data. Links are given to best practices and software tools for data management on the DataONE website; these links point to more in-depth descriptions, examples and rationale for the best practices and software tools. Although many of the best practices were created with tabular (i.e. spreadsheet) data in mind, many of the concepts are applicable to other types of data produced by scientists, including databases, images, gridded data, or shape files.

We provide a guide to data management practices that investigators could perform during the course of data collection, processing, and analysis (i.e. components of the data life cycle, Fig. 1) to improve the chances of their data being used effectively by others (reused). These practices could be performed at any time during the preparation of the data set, but we suggest that researchers consider them in the data management planning stage, before the first measurements are taken. In addition, sometimes steps of the life cycle (and data management in general) can and should occur simultaneously; for instance, describing your collection methods is easier during the collection phase, rather than trying to reconstruct methods later to add to your data documentation.

4. The Data Life Cycle: An Overview

The data life cycle has eight components:

Plan: description of the data that will be compiled, and how the data will be managed and made accessible throughout its lifetime
Collect: observations are made either by hand or with sensors or other instruments and the data are placed a into digital form
Assure: the quality of the data are assured through checks and inspections
Describe: data are accurately and thoroughly described using the appropriate metadata standards
Preserve: data are submitted to an appropriate long-term archive (i.e. data center)
Discover: potentially useful data are located and obtained, along with the relevant information about the data (metadata)
Integrate: data from disparate sources are combined to form one homogeneous set of data that can be readily analyzed
Analyze: data are analyzed

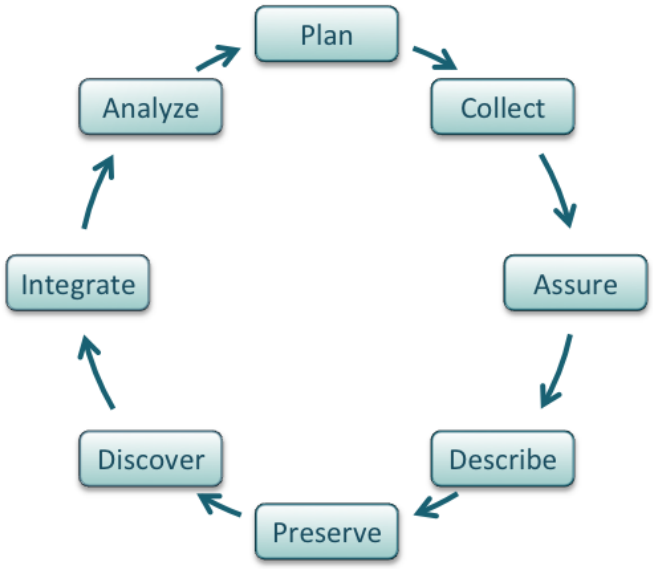

Figure 1: The data life cycle from the perspective of a researcher. The 'plan' component describes the entire life cycle.

Some projects might use only part of the life cycle; for instance, a project involving meta-analysis might focus on the Discover, Integrate, and Analyze steps, while a project focused on primary data collection and analysis might bypass the Discover and Integrate steps. In addition, other projects might not follow the linear path depicted here, or multiple revolutions of the cycle might be necessary.

A scientist or team of scientists is frequently engaged in all aspects of the data life cycle, both as a data creator and as a data user. Some scientists or teams (e.g. those engaged in modeling and synthesis) may create new data in the process of discovering, integrating, analyzing, and synthesizing existing data. This primer summarizes best practices [8, 9, 10] for preparing data that can be readily shared with others.

5. Data Management Throughout the Data Life Cycle

5.1. Plan

Plan for data management as your research proposal (for funding agency, dissertation committee, etc.) is being developed. Revisit your data management plan frequently during the project and make changes as necessary. Consider the following:

- Collecting your data: Based on the hypotheses and sampling plan, what data will be generated? How will the samples be collected and analyzed? Provide descriptive documentation of collection rationale and methods, analysis methods, and any relevant contextual information. What instruments will be used? What cyberinfrastructure will be required to support your research?
- Decide on a repository: Select a data repository (i.e. data center) that is most appropriate for the data you will generate and for the community that will make use of the data. Talk with colleagues and research sponsors about the best repository for your discipline and your type of data. Check with the repository about requirements for submission, including required data documentation, metadata standards, and any possible restrictions on use(e.g. intellectual property rights). Repository specifications will help guide decisions about the remainder of the practices below.
- Organizing your data: Decide on how data will be organized within a file, what file formats will be used, and the types of data products you will generate. Does your community have standard formats (file formats, units, parameter names)? Consider whether a relational database or other data organization strategy might be most appropriate for your research.
- Managing your data: Who is in charge of managing the data? How will version control be handled? How will data be backed up, and how often?
- Describing your data: How will you produce a metadata record? Using what metadata standard? Using what tool? Will you create a record at the project's inception and update it as you progress with your research? Where will you deposit the metadata? Consider your community's standards when deciding on the metadata standard and data center.
- Sharing your data: Develop a plan for sharing data with the project team, with other collaborators, and with the broader science community. Under what conditions will data be released to each of these groups? What are the target dates for release to these groups? How will the data be released?
- Preserving your data: As files are created, implement a short-term data preservation plan that ensures that data can be recovered in the event of file loss (e.g., storing data routinely in different locations).
- Consider your budget: What types of personnel will be required to carry out your data management plan? What types of hardware, software, or other computational resources will be needed? What other expenses might be necessary, such as data center donation or payment? The budget prepared for your research project should include estimated costs for data management.
- Explore your institutional resources: Some institutions have data management plan templates, suggested institutional data centers, budget suggestions, and useful tools for planning your project.

5.2. Collect

It is important to collect data in such a way as to ensure its usability later. Careful consideration of methods and documentation before collection occurs is important.

- Consider creating a template for use during data collection. This will ensure that any relevant contextual data are collected, especially if there are multiple data collectors.
- Describe the contents of your data files: Define each parameter, including its format, the units used, and codes for missing values (see here). Provide examples of formats for common parameters. Data descriptions should accompany the data files as a “readme.txt” file, a metadata file using an accepted metadata standard, or both.

- Use consistent data organization: We recommend that you organize the data within a file in one of the two ways described below. Whichever style you use, be sure to place each observation on a separate line (row).
    1. Each row in a file represents a complete record and the columns represent all the parameters that make up the record (a spreadsheet format).
    2. One column is used to define the parameter and another column is used for the value of the parameter (a database format). Other columns may be used for site, date, treatment, units of measure, etc. For specific examples, refer to [10].
- Use the same format throughout the file; for instance, do not rearrange columns or rows within the file. At the top of the file, include one or more header rows that identify the parameter and the units for each column. "Atomize" data: make sure there is only one piece of data in each entry.
- Use plain text ascii characters for variable names, file names, and data: this will ensure that your data file is readable by the maximum number of software programs.
- Use stable, non-proprietary software and hardware: File formats should ideally be non-proprietary (e.g. .txt or .csv files rather than .xls), so that they are stable and can be read well into the future. Consider the longevity of hardware when backing up data.
- Assign descriptive file names: File names ideally describe the project, file contents, location, and date, and should be unique enough to stand alone as file descriptions. File names do not replace complete metadata records.
- Keep your raw data raw: Preserve the raw data, with all of its imperfections. Use a scripted program to "clean" the data so that all steps are documented.
- Create a parameter table: Describe the code and abbreviations used for a parameter, the units, maximum and minimum values, the type of data (i.e. text, numerical), and a description.
- Create a site table: Describe the sites where data were collected, including latitude, longitude, dates visited, and any contextual details (e.g. ecosystem type, land cover or use, weather conditions, etc.) that might affect the data collected.

5.3. Assure

Perform basic quality assurance and quality control on your data (see here), during data collection, entry, and analysis. Describe any conditions during collection that might affect the quality of the data. Identify values that are estimated, double-check data that are entered by hand (preferably entered by more than one person), and use quality level flags (see here) to indicate potential problems. Check the format of the data to be sure it is consistent across the data set. Perform statistical and graphical summaries (e.g. max/min, average, range) to check for questionable or impossible values and to identify outliers. Communicate data quality using either coding within the data set that indicates quality, or in the metadata or data documentation. Identify missing values. Check data using similar data sets to identify potential problems. Additional problems with the data may also be identified during analysis and interpretation of the data prior to manuscript preparation.

5.4. Describe: Data Documentation

Comprehensive data documentation (i.e. metadata) is the key to future understanding of data. Without a thorough description of the context of the data file, the context in which the data were collected, the measurements that were made, and the quality of the data, it is unlikely that the data can be easily discovered, understood, or effectively used. Consider the following when documenting your data:

- Describe the digital context
    - Name of the data set
    - The name(s) of the data file(s) in the data set
    - Date the data set was last modified

        - Example data file records for each data type file
        - Pertinent companion files
        - List of related or ancillary data sets
        - Software (including version number) used to prepare/read the data set
        - Data processing that was performed
    - Describe the personnel and stakeholders
        - Who collected the data
        - Who should be contacted with questions
        - Sponsors
    - Describe the scientific context
        - Scientific reason why the data were collected
        - What data were collected
        - What instruments (including model and serial number) were used
        - Environmental conditions during collection
        - Where collected and spatial resolution
        - When collected and temporal resolution
        - Standards or calibrations used
    - Information about parameters
        - How each was measured or produced
        - Units of measure
        - Format used in the data set
        - Precision, accuracy, and uncertainty
        - Information about data
        - Taxonomic details
        - Definitions of codes used
        - Quality assurance and activities
        - Known problems that limit data use (e.g. uncertainty, sampling problems)
        - How to cite the data set

Metadata should be generated in a metadata format commonly used by the most relevant science community. Use metadata editing tools (e.g. Metavist [11], Mercury Metadata Editor [12], Morpho [13]) to generate comprehensive descriptions of the data. Comprehensive metadata enables others to discover, understand, and use your data.

5.5. Preserve

Work with a data center or archiving service that is familiar with your area of research. They can provide guidance as to how to prepare formal metadata, how to preserve the data, what file formats to use, and how to provide additional services to future users of your data. Data centers can provide tools that support data discovery, access, and dissemination of data in response to users needs.

- Identify data with long-term value: It is not necessary to archive all of the data products generated from your research. Consider the size of your files, which data will be most useful for future data users (typically raw data), and which data versions would be most difficult to reproduce.
- Store data using appropriate precision (e.g. significant digits)
- Use standard terminology: To enable others to find and use your data, carefully select terminology to describe your data. Use common keywords and consult ontologys for your discipline, if they are available.

- Consider legal and other policies: All research requires the sharing of information and data. Researchers should be aware of legal and policy considerations that affect the use and reuse of their data. It is important to provide the most comprehensive access possible with the fewest barriers or restrictions. There are three primary areas that need to be addressed when producing sharable data:
    1. Check your institution's policies on privacy and confidentiality.
    2. Data are not copyrightable. Ensure that you have the appropriate permissions when using data that has multiple owners or copyright layers. Information documenting the context of data collection may be under copyright.
    3. Data is able to be licensed. The manner in which you license your data can determine its ability to be consumed by other scholars. For example, the Creative Commons Zero License [14] provides for very broad access.

  If your data fall into any of the following categories, there are additional considerations regarding sharing: Rare, threatened or endangered species; Cultural items returned to their country of origin; Native American and Native Hawaiian human remains and objects; Any research involving human subjects. If you use data from other sources, you should review your rights to use the data and be sure you have the appropriate licenses and permissions.
- Attribution and provenance: The following information should be included in the data documentation or the companion metadata file:
    - The personnel responsible for the data set throughout the lifetime of the data set
    - The context of the data set with respect to a larger project or study (including links and related documentation), if applicable
    - Revision history, including additions of new data and error corrections
    - Links to source data, if the data were derived from another data set
    - Project support (e.g., funding agencies, collaborators, computational support)
    - How to properly cite the data set
    - Intellectual property rights and other licensing considerations

5.6. Discover, Integrate, and Analyze

A variety of tools are available that support data discovery, integration, analysis, and visualization. Significant recent advances have been made in supporting the creation and management of complex, scientific workflows that serve to integrate, analyze, and visualize data as well as document the exact steps used in those processes [15, 16, 17].

When data sets and data elements are used as a source for new data sets, it is important to identify and document those data within the documentation of the new derived data set (i.e. data set provenance). This will enable 1) tracing the use of data sets and data elements, 2) attribution to the creators of the original data sets, and 3) identifying effects of errors in the original data sets or elements of those sets on derived data sets.

6. Conclusion

Data represent important products of the scientific enterprise that are, in many cases, of equivalent or greater value than the publications that are originally derived from the research process. For example, addressing many of the grand challenge scientific questions increasingly requires collaborative research and the reuse, integration, and synthesis of data. Consequently, academic, research and funding institutions are now requiring that scientists provide good stewardship of the data they collect. By implementing good data management practices early in the data life cycle (Fig. 1), scientists can ensure that they are well prepared to meet these requirements.

The DataONE Best Practices database [18] represents an initial effort to educate scientists about best practices they can follow in managing their data. The database and accompanying primer (this paper) will continue to be updated in response to community feedback, as well as the availability of new enabling technologies. Further creation and refinement of educational resources such as the DataONE resources website and this primer are important for enabling good data stewardship. However, these products represent just one facet of the comprehensive education effort that is needed. In particular, we encourage professional societies to include data and information management training as a routine part of societal meetings because of the constant change in technology and the evolving expectations of research sponsors and the public. More importantly, we recommend that data management best practices be incorporated in introductory biology, ecology, and environmental science courses, and that stand-alone graduate courses on data management be included in the required curricula for graduate schools. Such sociocultural changes are necessary if the next generation of scientists is to be equally knowledgeable of current scientific information as well as the data and informatics practices that lead to information and knowledge.

7. Acknowledgements

This work was supported by the National Science Foundation (grant numbers 0753138 and 0830944) and the National Aeronautic and Space Administration (grant number NNG09HP121). We thank the many individuals who participated in the two workshops; their names and institutions are listed at www.dataone.org/best-practices.

8. References

[1] Piwowar, HA, RS Day, and DB Fridsma, Sharing Detailed Research Data Is Associated with Increased Citation Rate, PLoS ONE 2(3): e308. 2007, DOI: 10.1371/journal.pone.0000308

[2] Michener, WK, JW Brunt, JJ Helly, TB Kirchner, and SG Stafford, Nongeospatial metadata for the ecological sciences, Ecological Applications 7(1):330-342, 1997.

[3] Committee on Geophysical Data of the NRC Commission on Geosciences, Solving the Global Change Puzzle: A U.S. Strategy for Managing Data and Information, The National Academies Press, Washington D.C. 1991.

[4] Data Management and Sharing Policies for Federal Funding Agencies, compiled by UC3, Accessed June 2011. http://www.cdlib.org/services/uc3/datamanagement/funding.html

[5] Joint Data Archiving Policy, Dryad, Accessed June 2011. http://datadryad.org/jdap

[6] US Government Accountability Office, United States Government Accountability Office 2007 Report. US Government Accountability Office, 2007. www.gao.gov/products/GAO-07-1172.

[7] National Science Foundation, Summary of Dissemination and Sharing of Research Results, Accessed June 2011. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[8] Cook, RB, RJ Olson, P Kanciruk, and LA Hook, Best practices for preparing ecological data sets to share and archive, Bulletin of the Ecological Society of America 82(2): 138-141. 2001

[9] Borer, ET, EW Seabloom, MB Jones, and M Schildhauer, Some simple guidelines for effective data management, Bulletin of the Ecological Society of America April 2009: 205-214.

[10] LA Hook, S Santhana-Vannan, TW Beaty, RB Cook, and BE Wilson, Best Practices for Preparing Environmental Data Sets to Share and Archive, Accessed June 2011. http://daac.ornl.gov/PI/BestPractices-2010.pdf

[11] D Rug, Metavist Metadata Editor, Accessed June 2011. http://metavist2.codeplex.com.

[12] Oak Ridge National Laboratory, Mercury Metadata Management, Data Discovery, and Access System, Accessed June 2011. http://mercury.ornl.gov

[13] Knowledge Network for Biocomplexity, Data Management Software Portal, Accessed June 2011. http://knb.ecoinformatics.org/morphoportal.jsp

[14] Creative Commons, CC0 License, Accessed June 2011. http://creativecommons.org/choose/zero/

[15] Ludaescher, B, I Altintas, C Berkley, D Higgins, E Jaeger-Frank, M Jones, E Lee, J Tao, and Y Zhao, Scientific Workflow Management and the Kepler System, Concurrency and Computation: Practice and Experience 18(10). 2006.

[16] Oinn, T, M. Greenwood, M Addis, J Ferris, K Glover, C Goble, D Hull, D Marvin, P. Li, and P Lord, Taverna: Lessons in creating a workflow environment for the life sciences, Concurrency and Computation: Practice and Experience 18(10):1067-1100. 2006.

[17] Goble, C and D DeRoure, \Part 3: Scientific Infrastructure" from The 4th Paradigm. Microsoft Research, Redmond, WA. 2009.

[18] DataONE Best Practices, DataONE online resources for Best Practices and Tools for Data Management, Accessed June 2011. https://www.dataone.org/best-practices.

9. Glossary

access (1) Mechanisms for obtaining or retrieving data or information from an instrument, sensor network, storage device, or data center or (2) Rights to download or use data.

archive To place or store data in a data center; typically done for ensuring long-term preservation of the data and to promote discovery and use of the data.

ascii A character-encoding scheme based on alphabet order, used to represent text in computers.

attribution Acknowledgment of the role that an individual, group, institution, or research sponsor played in support of a research project and the resulting products (e.g. papers and data).

best practice Methods or approaches that are typically recognized by a community as being correct or most appropriate for acquiring, managing, analyzing, and sharing data.

coding Creating a program that can be understood and acted upon by a computer.

companion metadata Data documentation that accompanies a data.

cyberinfrastructure Structure that consists of systems for computing and data storage, repositories, and computing tools that are linked by networks, providing more powerful capabilities discovery and innovation. product (e.g. textual descriptions of the rows and columns of a data table as well as the scientific context underlying the data).

data center A facility that contains computers and data storage devices and that is used for data storage and transmission (e.g., acquiring data from providers and making data available to users). Data centers frequently provide curation and stewardship services, access to data products, user help desk support and training, and sometimes support data processing activities and other value-added services.

data documentation The metadata or information about a data product (e.g., data table, database) that enables one to understand and use the data. Such information may include the scientific context

underlying the data as well as who collected the data, why the data were collected, and where, when, and how the data were collected.

data entropy Normal degradation in information content associated with data and metadata over time (paraphrased from [2]).

database An organized collection of data. A database can be classified by the type of content included in it (e.g., bibliographic, statistical, document-text) or by its application area (e.g., Biological, Geological, etc).

deposit The act of submitting data, as to a repository.

derived data set A new dataset created by using multiple existing datasets and their data elements as sources. Also refers to a new dataset created by the addition of a single dataset, used as a source with newly collected data.

discover The act of finding new data.

dissemination The act of spreading widely. Data dissemination refers to making data available from one or multiple sources.

file format The specific organization of information in a digital computer file.

header row A meaningful name for referencing the content contained in a row or column, as in a spreadsheet.

impossible value An unreasonable value, outside the working range for a parameter, that should be identified in the QA/QC process; for example an air temperature of 190C for the Earth or a pH of -2 are impossible and should be tagged as such.

meta-analysis An analysis that combines the results of many studies.

metadata Data that provides descriptive information (content, context, quality, structure, and accessibility) about a data product and enables others to search for and use the data product.

metadata editing tool A software tool to input, edit, and view metadata; output from the tool is metadata in a standard extensible markup language (xml) format.

metadata format Standardized structure and consistent content for metadata, usually in machine readable extensible markup language (xml) that can be represented in other human readable formats (e.g., html, pdf, etc.). Standards for metadata include EML, FGDC, and ISO 19115.

metadata standard Requirements for metadata documentation that are intended ensure correct use and interpretation of the data by its owners and users. Different scientific communities use different sets of metadata standards; common examples are EML (Ecological Metadata language), FGDC (Federal Geographic Data Committee) standard, and ISO 19115 (International Organization for Standardization Geographic information metadata).

missing value A value that is not in the data file, because the information / sample was not collected, lost, not analyzed, or an impossible value, etc. A missing value code indicates that a value is missing and a missing value flag (a categorical parameter) describes the reason that the value was missing.

non-proprietary In the public domain, not protected by patent, copyright, or trademark.

ontology A framework for interrelated concepts within a domain; an ontology would link the terms "water vapor", "relative humidity", and "H_2O vapor pressure", so that a user searching for one, would also see the other related terms and their relationships. plural.

parameter A variable and measurable factor that determines or characterizes a system.

preserve Format and document data for long term storage and potential use.

provenance History of a data file or data set, including collection, transformations, quality control, analyses, or editing.

quality assurance A set of activities to ensure that the data are generated and compiled in a way that meets the goal of the project.

quality control Testing or other activities designed to identify problems in the data.

quality level flag An indicator within the data file that identifies the level of quality of a particular data point or data set. Flags can be defined within the metadata.

relational database A collection of tables (often called relations) with defined relationships to one another.

reuse Using data for a purpose other than that for which it was collected.

scientific workflow A precise description of scientific procedure, often conceptualized as a series of data ingestion, transformation, and analytical steps.

scripted program A program (requiring a command line interface or similar) that performs an action or task. Scripts can be saved, modified and re-used as necessary.

stable Something that is unlikely to become obsolete or undergo significant change due to changes in version development, funding etc.

stewardship The act of caring for, preserving or improving over time.

tool A device or implement (in this instance a piece of software) that is used to carry out an action or series of actions.

version control The task of keeping a software system consisting of many versions and configurations well organized.