DataONE Data Management Training - the different approaches to data management training

Rebecca Koskela
Belmont Forum Digital Skills and Curricula Development Workshop
April 28, 2017
Development of Resources

Approach

- Solicit community feedback
- Build partnerships
- Append, create, revise offerings
Challenges

Educators

Barriers to Teaching Data Management (n=134)

- There is no time to teach data management (n=69) - 51.5%
- It is not my area of expertise (n=53) - 39.6%
- I don't have enough information (n=41) - 30.6%
Sustainable Outreach
Online Engagement
Best Practices
Database and Primer

Best Practices

The DataONE Best Practices database provides guidelines on how to effectively work with your data through all stages of the Data Lifecycle. Users can access best practices within the database by either clicking on a stage of the lifecycle or selecting keywords under search.

Best Practices Primer
For students and others new to data management, we provide a Best Practices Primer as an introduction to the DataONE Best Practices database and data management in general.

Public Participation in Science Research Data Management Guide
We also provide a Data Management Guide written specifically for the Citizen Science community that takes the user through the stages of the data lifecycle and links to various DataONE Best Practices online.

Search Best Practices

Search by Keyword in title
Search by Keyword in Body

Filter by tag
access, analyze, annotation, assure, backup, calibration, citation, coding, collect

Filter by Data Life Cycle
- Any -

You may enter multiple tags by holding down command (control) and making your selection

Search Reset

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

* but were afraid to ask

Carly Stresser, Robert Cook, William Michener, Amber Budden

Contents
1. Objective of This Primer
2. Why Manage Data?
   2.1. It will benefit you and your collaborators
   2.2. It will benefit the scientific community
   2.3. Journals and sponsors want you to share your data
3. How To Use This Primer
4. The Data Life Cycle: An Overview
5. Data Management Throughout the Data Life Cycle
   5.1. Plan
   5.2. Collect
   5.3. Assure
   5.4. Describe: Data Documentation
   5.5. Preserve
   5.6. Discover, Integrate, and Analyze
6. Conclusion
7. Acknowledgements
8. References
9. Glossary

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].

DataONE Best Practices Primer

www.dataone.org/best-practices
Best Practices

Metrics

**Primer**

3500+ unique downloads
(most downloaded resource)

**Web page**

5800+ unique visits
(4th most visited page, 2nd most visited area)
Data Management Modules

Overview

Lesson 10: Analysis and Workflows

Typical data analyses
- Data processing: may include selecting a subset of data for analysis, merging multiple data sets, manipulating data for suitability, or data transformation.
- Graphical analysis: makes it easier to see patterns and can aid in the identification of outliers.
- Statistical analysis: conventional statistics are used to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data.

Science is iterative: the process that results in a scientific product can be complex.

Reproducibility...
- Is at the core of the scientific process. If results are not reproducible, they lose credibility.
- Good documentation of the data used in the analysis is essential.

Workflows
- Informal workflow: no special software is needed to create work diagrams. Workflow diagrams include:
  - Inputs and outputs
  - Transformation rules or analytical processes
  - Decision points
  - Arrows indicating direction of process flow

- Formal workflow: analytical pipeline where each step can be implemented in different software packages. Parameters and requirements for each step are recorded.
  - Single access point for multiple analyses across software packages
  - Keeps track of analysis and provenance to better enable reproducibility
  - Workflow can be stored

Hands-on Activity 1: Accessing Data in the Literature

Lesson: 10: Why Data Management

Objectives: Students recognize the value of accessible archived data, by experiencing the challenges of accessing data from published papers.

Outcomes: [1] Students can explain why accessible data archiving is valuable. [2] Students can provide strategies for getting data from published papers, and anticipate challenges to accessing the data.

Time Needed: One hour out of class, 15–30 minutes in class discussion

Links: Any resource for searching scientific literature (e.g., Web of Science, Google Scholar, DataONE, etc.)

Additional Files Needed: None


Notes and Instructions for Instructors:

An intended take-home lesson of this activity is that access to valuable original data can become difficult or impossible in a short period of time after a paper is published, but this loss of accessibility is avoidable. It is now easy to access original data depending on the field, some fields have developed a culture of data sharing and data accessibility, including genetics, climate studies, and geography. Others do not have this tradition. Because of these field-specific cultures, students’ success at accessing data will depend on the topic and question they choose.

It may be worth reviewing with the students the different ways by which scientists access other’s data: data tables or published data appendices within a paper, extracting (estimating) data from published graphs, online data archives or data streams (often restricted to journal subscribers or public). Writing the author and requesting the data.

After students have completed the exercise (see Student Instructions, below), have students discuss the challenges that they faced in figuring out how to access data from the published literature that are relevant to their question, and ways the students came up with to deal with the challenges. This can be done as a 15 to 30 minute whole class discussion or in small groups. Things to note include: whether accessibility to data varied depending on the question addressed, and whether accessibility depended on how long ago the work was published. Perhaps culminate the discussion with questions about why data underpinning...
Data Management Modules
Enhancement

- Comprehensive peer-review and revision
- Transition to GitHub