Educating Scientists about the Data Life Cycle

Bill Michener

Professor and DataONE Project Director
University of New Mexico

9 October 2012

2012 eScience Workshop
DataONE

Three major components for a flexible, scalable, sustainable network

<table>
<thead>
<tr>
<th>Member Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• diverse institutions</td>
</tr>
<tr>
<td>• serve local community</td>
</tr>
<tr>
<td>• provide resources for managing their data</td>
</tr>
<tr>
<td>• retain copies of data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coordinating Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• retain complete metadata catalog</td>
</tr>
<tr>
<td>• indexing for search network</td>
</tr>
<tr>
<td>• network-wide services</td>
</tr>
<tr>
<td>• ensure content availability (preservation)</td>
</tr>
<tr>
<td>• replication services</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Investigator Toolkit</th>
</tr>
</thead>
<tbody>
<tr>
<td>[List of tools and services]</td>
</tr>
</tbody>
</table>
The Data Life Cycle

- Plan
- Collect
- Describe
- Assure
- Preserve
- Discover
- Integrate
- Analyze
User Assessments

Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir¹*, Suzie Allard¹, Kimberly Douglass¹, Arsev Umur Aydinoglu¹, Lei Wu¹, Eleanor Read², Maribeth Manoff², Mike Frame³

¹ School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, ² University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, ³ Center for Biological Informatics, United States Geological Survey, Oak Ridge, Tennessee, United States of America

Abstract

Background: Scientific research in the 21st century is more data intensive and collaborative than in the past. It is important to study the data practices of researchers – data accessibility, discovery, re-use, preservation and, particularly, data sharing. Data sharing is a valuable part of the scientific method allowing for verification of results and extending research from prior results.

Methodology/Principal Findings: A total of 1329 scientists participated in this survey exploring current data sharing practices and perceptions of the barriers and enablers of data sharing. Scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding. Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle (collecting their research data; searching for, describing or cataloging, analyzing, and short-term storage of their data) but are not satisfied with long-term data preservation. Many organizations do not provide support to their researchers for data management both in the short-
Education

- Best Practices
- Software Tools Catalog
- In-depth Training
Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle (shown below). Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

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.

The development of the DataONE Best Practices database was a collaborative effort across many individuals (credits).
Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle (shown below). Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

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.

The development of the DataONE Best Practices database was a collaborative effort across many individuals (credits).

Data Life Cycle

- Plan
- Collect
- Assure
- Describe
- Preserve
- Integrate
- Discover
- Analyze
- Plan
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 understand the context in which it was generated? Can they use it for their own research? Data management is the process of ensuring that this happens.

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

How To Use This Primer

The Data Life Cycle: An Overview

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

Conclusion

Acknowledgements

References

Glossary
Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle (shown below). Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

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.

The development of the DataONE Best Practices database was a collaborative effort across many individuals (credits).
Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle (shown below). Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

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.

The development of the DataONE Best Practices database was a collaborative effort across many individuals (credits).
**describe**

**Assign descriptive file names**

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type.

Read more

**Choose and use standard terminology to enable discovery**

Terms and phrases that are used to represent categorical data values or for creating content in metadata records should reflect appropriate and accepted vocabularies in your community or institution. Methods used to identify and select the proper terminology include:

Read more

**Confirm a match between data and their description in metadata**

To assure that metadata correctly describes what is actually in a data file, visual inspection or analysis should be done by someone not otherwise familiar with the data and its format. This will assure that the metadata is sufficient to describe the data. For example, statistical software can be used to summarize data contents to make sure that data types, ranges and, for categorical data, values found, are as described in the documentation/metadata.

Read more

**Create a data dictionary**

A data dictionary provides a detailed description for each element or variable in your dataset and data model. Data dictionaries are
**describe**

**Assign descriptive file names**

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type.

Read more

---

**Choose and use standard terminology to enable discovery**

Terms and phrases that are used to represent categorical data values or for creating content in metadata records should reflect appropriate and accepted vocabularies in your community or institution. Methods used to identify and select the proper terminology include:

Read more

---

**Confirm a match between data and their description in metadata**

To assure that metadata correctly describes what is actually in a data file, visual inspection or analysis should be done by someone not otherwise familiar with the data and its format. This will assure that the metadata is sufficient to describe the data. For example, statistical software can be used to summarize data contents to make sure that data types, ranges and, for categorical data, values found, are as described in the documentation/metadata.

Read more

---

**Create a data dictionary**

A data dictionary provides a detailed description for each element or variable in your dataset and data model. Data dictionaries are
Assign descriptive file names

Best Practice:

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type.

When choosing a file name, check for any database management limitations on file name length and use of special characters. Also, in general, lower-case names are less software and platform dependent. Avoid using spaces and special characters in file names, directory paths and field names. Automated processing, URLs and other systems often use spaces and special characters for parsing text string. Instead, consider using underscore (_) or dashes (-) to separate meaningful parts of file names. Avoid $ % ^ & # | : and similar.

If versioning is desired a date string within the file name is recommended to indicate the version.

Avoid using file names such as mydata.dat or 1998.dat.

Description Rationale:

Clear, descriptive, and unique file names may be important when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators. File names that reflect the contents of the file and uniquely identify the data file enable precise search and discovery of particular files.

Additional Information:


Examples:

An example of a good data file name:

Sevilleita_LTER_NM_2001_NPP.csv

Sevilleita_LTER is the project name
NM is the state abbreviation
2001 is the calendar year
NPP represents Net Primary Productivity data
csv stands for the file type—ASCII comma separated variable

Instead of "data May2011" use "data_May2011" or "data-May2011"

Tags: access format describe discover
Software Tools Catalog

The Software Tools database is the product of two NSF-funded Informatics Education Planning Workshops hosted by DataONE. The database provides a brief description of a wide range of tools that are recommended for use by scientists and students, as well as additional information and links to further resources. Users can access tools within the database by selecting keywords (under advanced search) or using free search.

The development of the DataONE Software Tools database was a collaborative effort across many individuals (credits).

View All Software Tools
Software Tools Catalog

The Software Tools database is the product of two NSF-funded Informatics Education Planning Workshops hosted by DataONE. The database provides a brief description of a wide range of tools that are recommended for use by scientists and students, as well as additional information and links to further resources. Users can access tools within the database by selecting keywords (under advanced search) or using free search.

The development of the DataONE Software Tools database was a collaborative effort across many individuals (credits).

View All Software Tools
visualization

3D World Studio

3D World Studio is a modeling program useful for visualizing real world data utilizing tools developed within computer gaming environments. The program allows you to create buildings and terrain and export your visualization into a variety of formats.

Adobe Illustrator

Professional standard software for creating original vector-based graphics. Includes powerful drawing tools and brushes.

The .ai file format is a common vector format for exchange and its feature set allows creation of complex vector artwork. Illustrator imports over two dozen formats (including PDF and SVG). Of particular use to data visualization is importation of SVG, or scalar vector graphics, which is an \\o3C recommendation, and is often exported from other programs.

Adobe Photoshop

Photoshop is a comprehensive photo editing tool produced by Adobe Systems. Users can manipulate photos, graphics, and other raster images using a variety of tools and predefined filters. Photoshop also allows users to record specific photo editing steps, which allows for automated batch processing. Photoshop is available as a stand-alone product, but is also part of Adobe's "Creative Suite" family of products. Photoshop Extended is an enhancement to Photoshop, and provides for enhanced 3D creation and editing.

Amber

"Amber" refers to two things: a set of molecular mechanical force fields for the simulation of biomolecules (Amber) and a package of molecular simulation programs which includes source code and demos (AmberTools).

ArcGIS Desktop

ArcGIS Desktop is a collection of software products for building complete geographic information systems (GIS) produced by Esri. ArcGIS Desktop 9 provides an integrated GIS, combining object-oriented and traditional file-based data models with a set of tools
visualization

3D World Studio

3D World Studio is a modeling program useful for visualizing real world data utilizing tools developed within computer gaming environments. The program allows you to create buildings and terrain and export your visualization into a variety of formats.

Read more

Adobe Illustrator

Professional standard software for creating original vector-based graphics. Includes powerful drawing tools and brushes.

The ai file format is a common vector format for exchange and its feature set allows creation of complex vector artwork. Illustrator imports over two dozen formats (including PDF and SVG). Of particular use to data visualization is importation of SVG, or scalar vector graphics, which is an W3C recommendation, and is often exported from other programs.

Read more

Adobe Photoshop

Photoshop is a comprehensive photo editing tool produced by Adobe Systems. Users can manipulate photos, graphics, and other raster images using a variety of tools and predefined filters. Photoshop also allows users to record specific photo editing steps, which allows for automated batch processing. Photoshop is available as a stand-alone product, but is also part of Adobe’s “Creative Suite” family of products. Photoshop Extended is an enhancement to Photoshop, and provides for enhanced 3D creation and editing.

Read more

Amber

"Amber" refers to two things: a set of molecular mechanical force fields for the simulation of biomolecules (Amber) and a package of molecular simulation programs which includes source code and demos (AmberTools).

Read more

ArcGIS Desktop

ArcGIS Desktop is a collection of software products for building complete geographic information systems (GIS). produced by Esri. ArcGIS Desktop 9 provides an integrated GIS, combining object-oriented and traditional file-based data models with a set of tools...
ArcGIS Desktop

ArcGIS Desktop is a collection of software products for building complete geographic information systems (GIS) produced by Esri. ArcGIS Desktop 9 provides an integrated GIS, combining object-oriented and traditional file-based data models with a set of tools to create and work with geographic data. The following three applications comprise the ArcGIS Desktop software suite:

- **ArcMap (mapping and data manipulation):** ArcMap is a map-authoring application
- **ArcCatalog (data management):** shared ArcGIS application that allows you to organize and access all GIS information (e.g., maps, layers, datasets, models, metadata, and services). Includes tools for browsing and finding geographic information; recording, viewing, and managing metadata; viewing datasets; and defining the schema structure of the object-based geographic datasets.
- **ArcToolbox (data conversion, modeling, and spatial analysis):** includes tools to do geographic feature overlay, feature selection and analysis, topology processing, and data conversion resulting in an output dataset. The geoprocessing framework allows you to use each geoprocessing function in a variety of ways. The tools can be used directly from a dialog, executed via command line, combined with other processes in visual models using Model Builder, or used in advanced scripts.

**Technical Expertise Required:** Basic programming skills

**Cost:** Cost-basis

**Website for more information:** Esri - The GIS Software Leader

**Additional Information:**

**Platform:** Windows, UNIX

**Tags:** GIS, map, analyze, visualization
In-depth Training

Education Modules

Below are links to education modules in powerpoint format that you can download and incorporate into your teaching materials.

The topics covered include:

- Lesson 01: Why Data Management
- Lesson 02: Data Sharing
- Lesson 03: Data Management Planning
- Lesson 04: Data Entry and Manipulation
- Lesson 05: Data Quality Control and Assurance
- Lesson 06: Data Protection and Backups
- Lesson 07: Metadata
- Lesson 08: How to Write Good Quality Metadata
- Lesson 09: Data Citation
- Lesson 10: Analysis and Workflows

If you use or consider using these materials, we would be grateful if you would take the opportunity to provide feedback.

Credits: Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, Lynda Wayne
In-depth Training

Education Modules

Below are links to education modules in powerpoint format that you can download and incorporate into your teaching materials.

The topics covered include:

- Lesson 01: Why Data Management
- Lesson 02: Data Sharing
- Lesson 03: Data Management Planning
- Lesson 04: Data Entry and Manipulation
- Lesson 05: Data Quality Control and Assurance
- Lesson 06: Data Protection and Backups
- Lesson 07: Metadata
- Lesson 08: How to Write Good Quality Metadata
- Lesson 09: Data Citation
- Lesson 10: Analysis and Workflows

If you use or consider using these materials, we would be grateful if you would take the opportunity to provide feedback.

Credits: Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, Lynda Wayne
Tutorials on Data Management

Lesson 10: Analysis and Workflows

Credits: Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, and Linda Wayne
Lesson Topics

1. Review of typical data analyses
2. Reproducibility & provenance
3. Workflows in general
4. Informal workflows
5. Formal workflows
Learning Objectives

After completing this lesson, the participant will be able to:

- Understand a subset of typical analyses used
- Define a workflow
- Understand the concepts informal and formal workflows
- Discuss the benefits of workflows
The Analysis Education Module
7 Lessons from Evaluation of Modules*

1. Use concrete or ‘real-world’ examples and stories to illustrate important points
2. Include information about (and links to) tools and resources
3. Use text sparingly on slides
4. Define jargon
5. Take data management experience levels into account
6. Include information about best practices
7. For a workshop format remove redundant information

*May 23-24, 2012 – 2 day training and content evaluation workshop; Credits: Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, and Linda Wayne
Walter E. Dean Environmental Information Management Institute

June 3-21, 2013
University of New Mexico
Walter E. Dean Environmental Information Management Institute

- 6 graduate credits
- 3 weeks
- Intensive, hands-on training
  - DMP Tool
  - Excel, Powerpoint
  - R
  - MySQL
  - ArcGIS
  - Kepler
  - Web design and Drupal
In-depth Training

- Plan
- Collect
- Assure
- Describe
- Preserve
- Discover
- Integrate
- Analyze

Tools:
- DMP Tool
- Excel
- DataUp
- MATLAB
- Kepler
- ONER
- Semantic Tools
- Mercury
- Mendeley
- zotero
Credits (Best Practices, Software Tools, Education Modules, EIM Summer Institute)

**Best Practices and Software Tools:**

**EIM Summer Institute:**

1. Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, and Linda Wayne

DataONE Team and Sponsors

- Amber Budden, Roger Dahl, Rebecca Koskela, Bill Michener, Robert Nahf, Skye Roseboom, Mark Servilla
- Dave Vieglais
- Suzie Allard, Nick Dexter, Kimberly Douglass, Carol Tenopir, Robert Waltz, Bruce Wilson
- John Cobb, Bob Cook, Ranjeet Devarakonda, Giri Palanismy, Line Pouchard
- Patricia Cruse, John Kunze
- Sky Bristol, Mike Frame, Richard Huffine, Viv Hutchison, Jeff Morisette, Jake Weltzin, Lisa Zolly
- Stephanie Hampton, Chris Jones, Matt Jones, Ben Leinfelder, Andrew Pippin
- Paul Allen, Rick Bonney, Steve Kelling
- Ryan Scherle, Todd Vision
- Randy Butler

- Ewa Deelman
- Deborah McGuinness
- Jeff Horsburgh
- Robert Sandusky
- Bertram Ludaescher
- Peter Honeyman
- Cliff Duke
- Carole Goble
- Donald Hobern
- David DeRoure