Educating Scientists about the Data Life Cycle

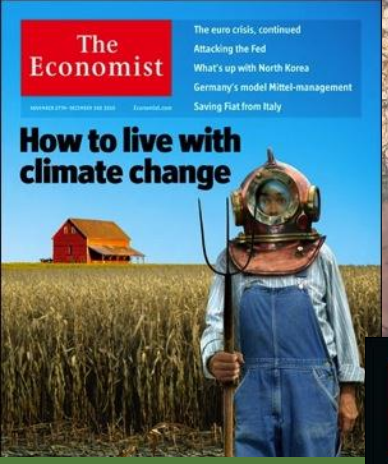
Bill Michener

Professor and DataONE Project Director
University of New Mexico

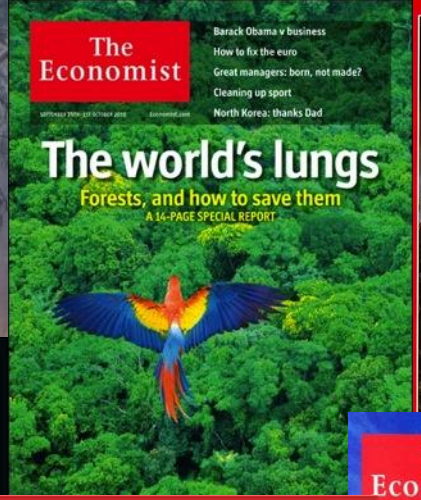
9 October 2012

2012 eScience Workshop

DataONE


The Economist
The euro crisis, continued
Attacking the Fed
What's up with North Korea
Germany's model Mittel-management
Saving Fiat from Italy
How to live with climate change


SCIENCE


The Economist
Barack Obama v business
How to fix the euro
Great managers: born, not made?
Cleaning up sport
North Korea: thanks Dad
The world's lungs
Forests, and how to save them
A 14-PAGE SPECIAL REPORT


HUD: Silent Sam Speaks
TIME
Torching the Amaz
Can the rain forest be saved?

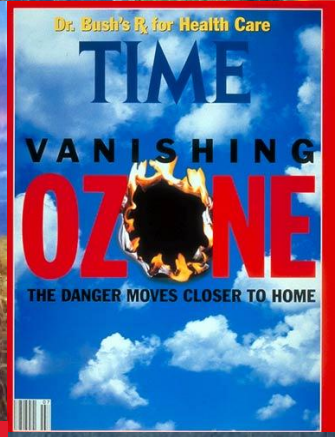

2ND-QTR SIZZLE
PROFITS AT 900 COMPANIES
PAYING FOR COLLEGE
BEWARE OF THOSE HIGH 529 FEES
TERRORISM
COMPANIES STILL NEED TO DO
BusinessWeek
GLOBAL WARMING
Why Business Is Taking It So Seriously
BY JOHN CAREY


The Economist
Silvio Berlusconi, your time is up
Iran throws down the gauntlet
Sovereign risk after Dubai
Has Obama got Afghanistan right?
Our books of the year
Stopping climate change
A 14-PAGE SPECIAL REPORT


Obama's Working-Class Woes
Why More Women Are Choosing C-Sections
Can Richard Branson Save the Airline Industry?
TIME
SPECIAL ENVIRONMENT ISSUE
How to Win The War On Global Warming
BY BRYAN WALSH


Living with Cancer
The changing science
Beyond Baghdad: Where The Enemy Has Its Own Surge
The Sopranos' Last Song: What Exit Will Tony Take?
TIME
SPECIAL DOUBLE ISSUE
The Global Warming Survival Guide
51 Things You Can Do to Make a Difference


The Economist
Brazil as the next oil giant
God help Italy
London's funny but sad election
The return of Disney
Cancer's link to stem cells
The silent tsunami
The food crisis and how to solve it


Dr. Bush's Rx for Health Care
TIME
VANISHING
OZONE
THE DANGER MOVES CLOSER TO HOME


China's Me Generation
Swinging Singles
Yep, He's Gay
TIME
Special Report
Why California Is Burning


TIME
Khomeini's Surprising Offer
Our Filthy Seas

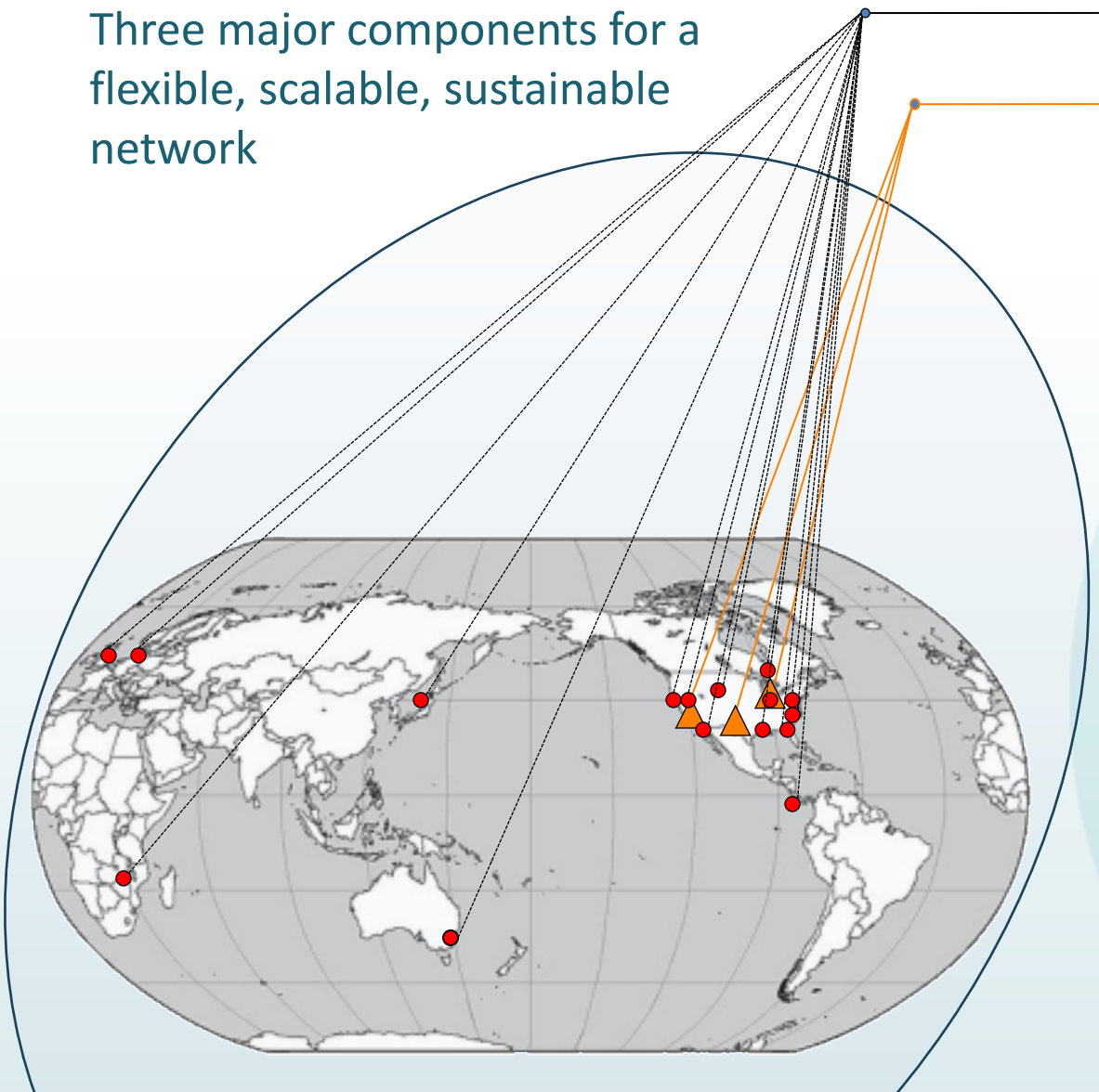

The Economist
East Asia's economies, five years on
Why Arab countries have failed
THE GLOBAL ENVIRONMENT
CO2AL
Environmental enemy No.1


SPECIAL REPORT
TIME
HOW TO SAVE THE EARTH
The hot and wild weather is a sign of things to come. But fresh ideas and new technology can cool us down and make this a GREEN CENTURY


Science
FORESTS IN FLUX
AAAS

DataONE

Three major components for a flexible, scalable, sustainable network



Member Nodes

Coordinating Nodes

Investigator Toolkit

MENDELEY RESEARCH NETWORKS

VEGBANK

Specify 6

Java

myexperiment

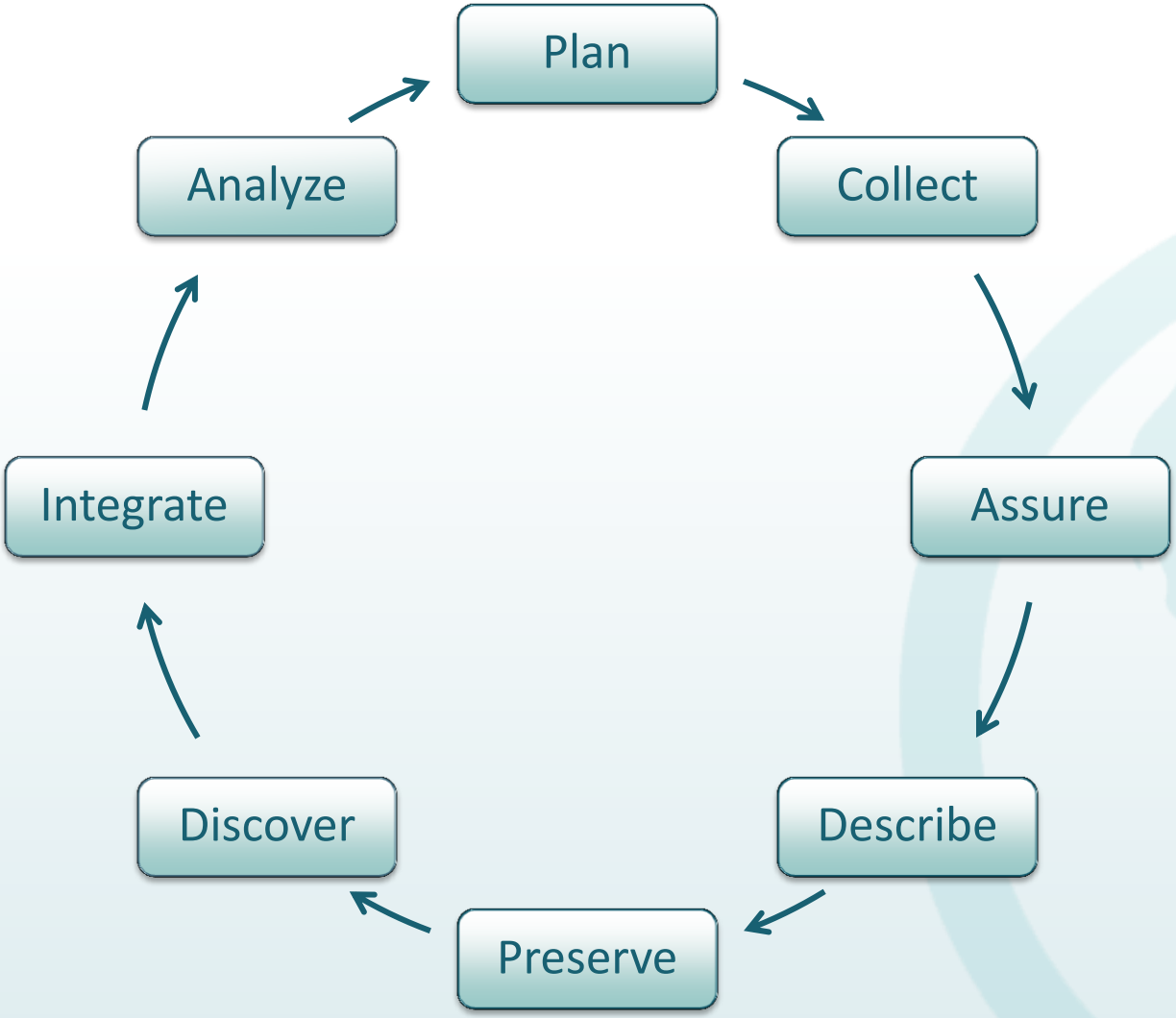
Excel:mac 2008

zotero

MATLAB The Language of Technical Computing

Kepler

The Data Life Cycle



User Assessments

OPEN ACCESS Freely available online

PLoS one

Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir[1]*, Suzie Allard[1], Kimberly Douglass[1], Arsev Umur Aydinoglu[1], Lei Wu[1], Eleanor Read[2], Maribeth Manoff[2], Mike Frame[3]

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

DataONE

Search DataONE Website For Go

About Participate Resources Education Data

Home » Resources » Best Practices

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance documentation metadata access assure plan describe format preserve analyze data archives quality

More

Featured Practice

Plan for effective multimedia management

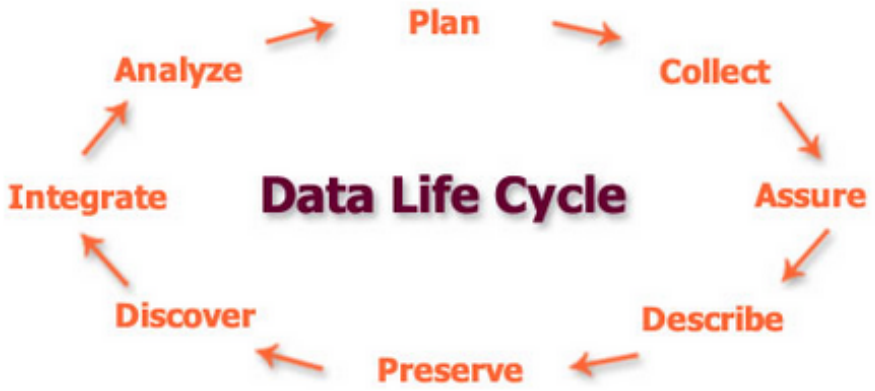
Multimedia data present

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).



View All Best Practices

Best Practices

DataONE

Search DataONE Website For Go

About Participate Resources Education Data

Home » Resources » Best Practices

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance documentation metadata access assure plan describe format preserve analyze data archives quality

More

Featured Practice

Plan for effective multimedia management

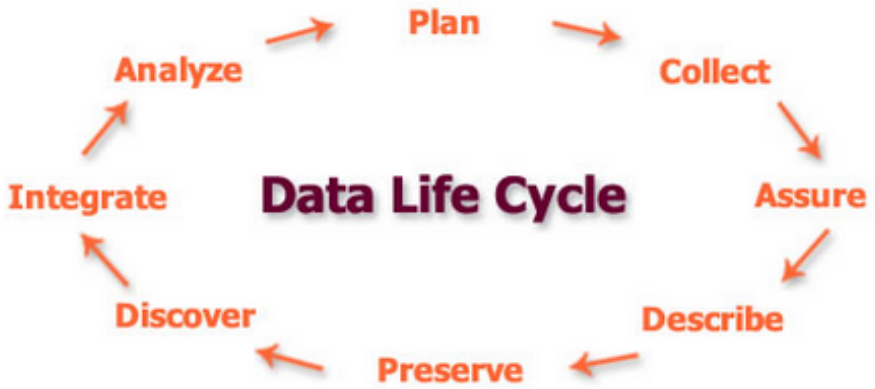
Multimedia data present

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** a[illegible] 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).



View All Best Practices

Best Practices Primer

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

Best Practices

DataONE

Search DataONE Website For Go

About Participate Resources Education Data

Home » Resources » Best Practices

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance documentation metadata access assure plan describe format preserve analyze data archives quality

More

Featured Practice

Plan for effective multimedia management

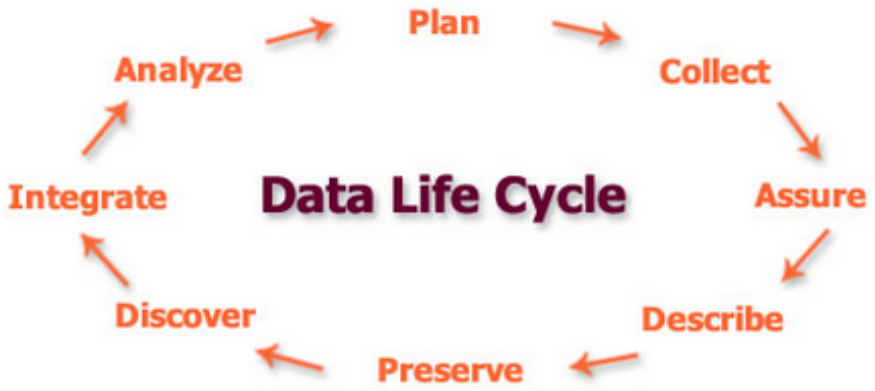
Multimedia data present

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).



View All Best Practices

Best Practices

DataONE

Search DataONE Website For Go

About Participate Resources Education Data

Home » Resources » Best Practices

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance documentation metadata access assure plan describe format preserve analyze data archives quality

More

Featured Practice

Plan for effective multimedia management

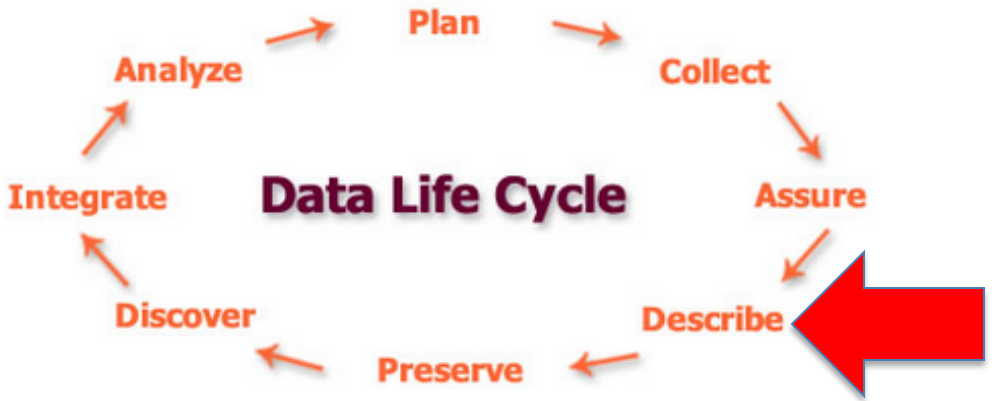
Multimedia data present

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).



View All Best Practices

DataONE

Search ONEMercury For Go

About Participate Resources Education Data

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance metadata preserve plan describe format quality analyze data archives access documentation assure

More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

Home » Resources » Best Practices » Taxonomy term

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

DataONE

Search ONEMercury For Go

About Participate Resources Education Data

Home » Resources » Best Practices » Taxonomy term

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

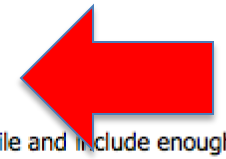
provenance metadata preserve plan describe format quality analyze data archives access documentation assure

More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

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

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

provenance plan
describe
documentation
data archives analyze
preserve access
assure metadata
format quality
More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

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:

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010

Borer et al. 2009. Some Simple Guidelines for Effective Data Management. Bull. of ESA 90: 209-214.

Examples:

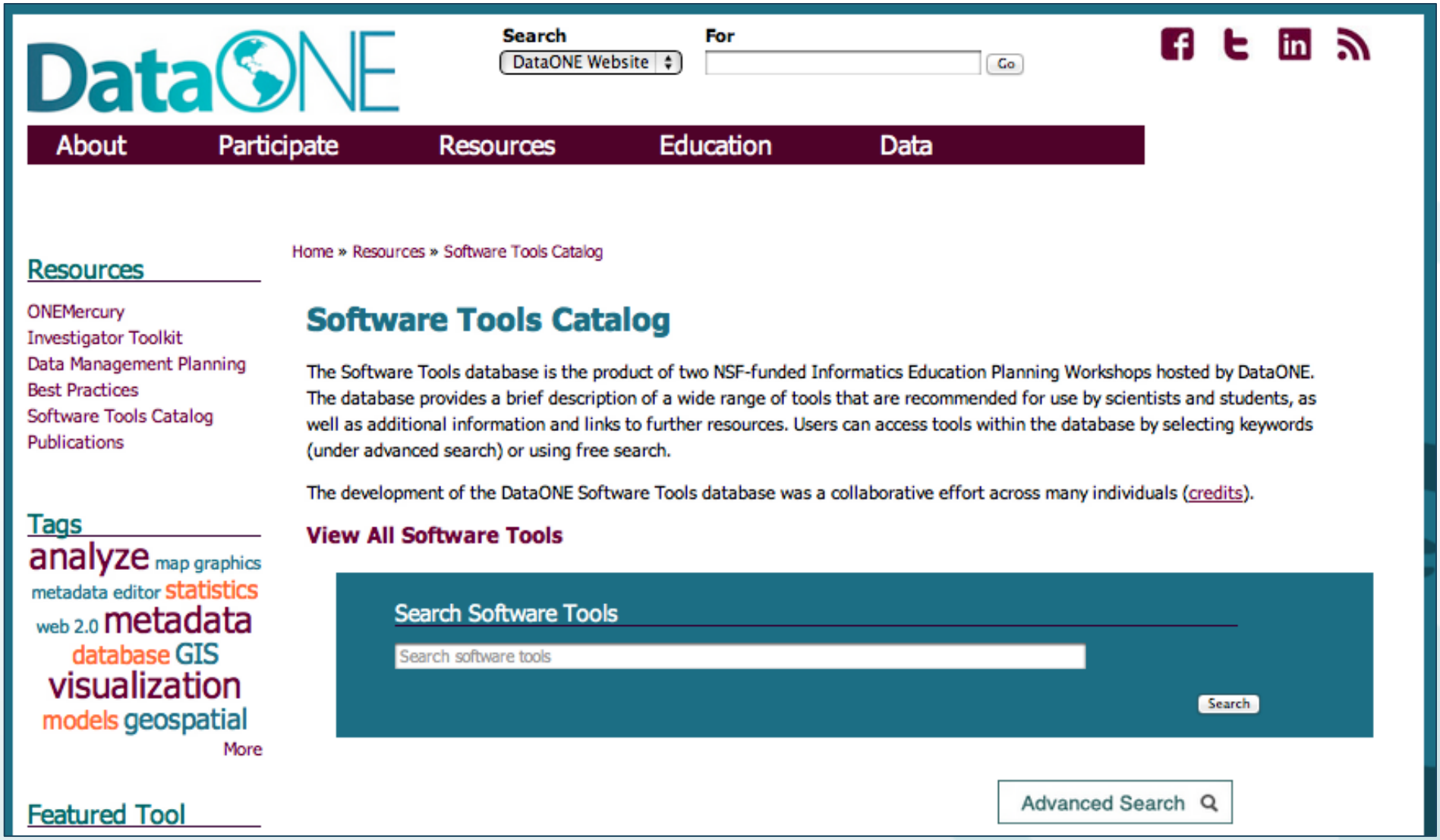
An example of a good data file name:

Sevilleta_LTER_NM_2001_NPP.csv
- Sevilleta_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

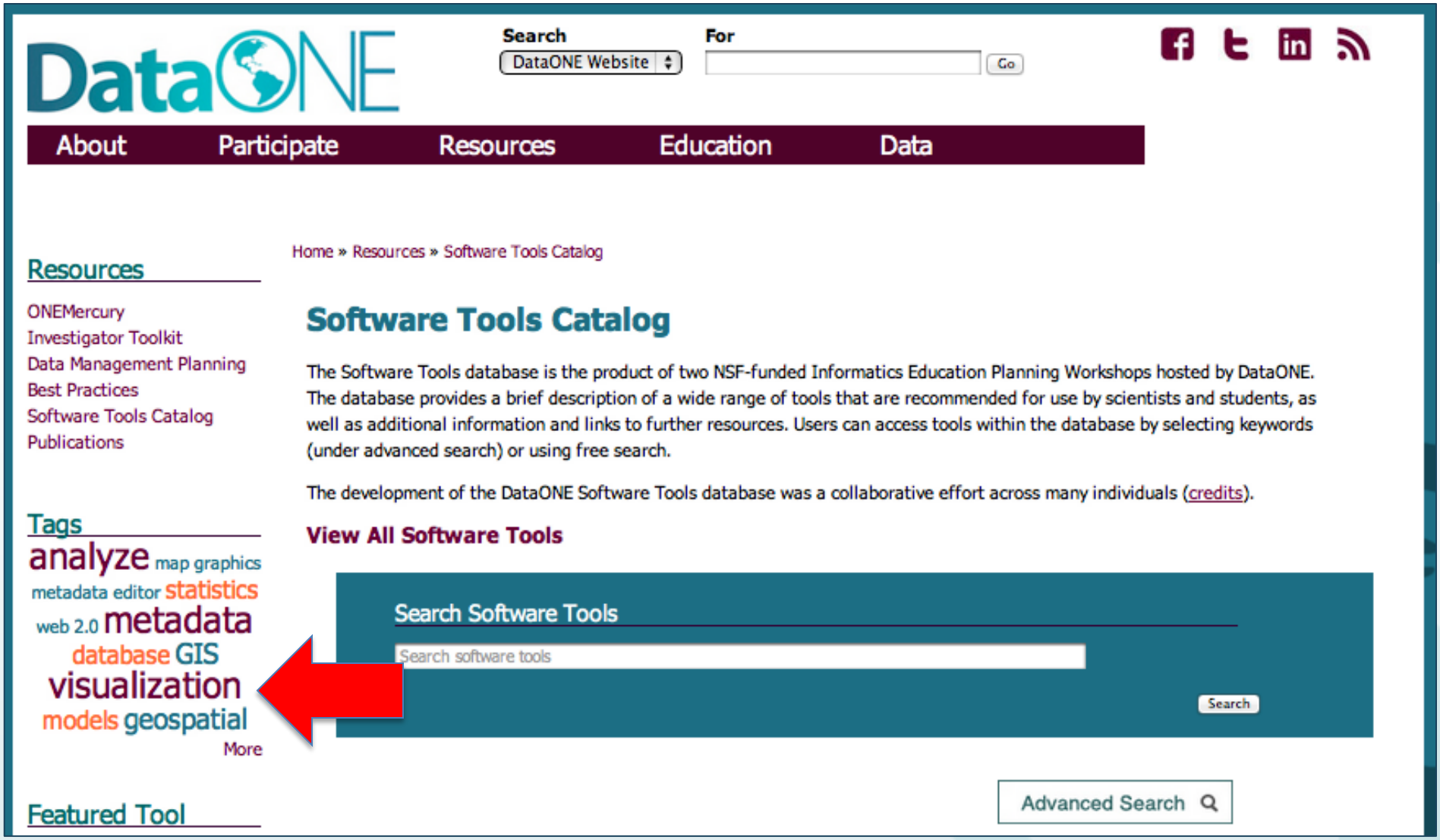
Instead of "data May2011" use "data_May2011" or "data-May2011"

Tags: access format describe discover

Software Tools Catalog



Software Tools Catalog



Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

visualization web 2.0 statistics map graphics analyze models database geospatial GIS metadata editor metadata

More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

Home » Resources » Software Tools Catalog » Taxonomy term

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

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

visualization web 2.0 statistics map graphics analyze models database geospatial GIS metadata editor metadata

More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

Home » Resources » Software Tools Catalog » Taxonomy term

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 [illegible]tware 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

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

graphics metadata editor
database statistics
geospatial
visualization
analyze map
models web 2.0 GIS
metadata
More

Find it Fast

Data
Data Management Planning
Best Practices
Software Tools

Home » Resources » Software Tools Catalog » ArcGIS Desktop

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, globes, 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.

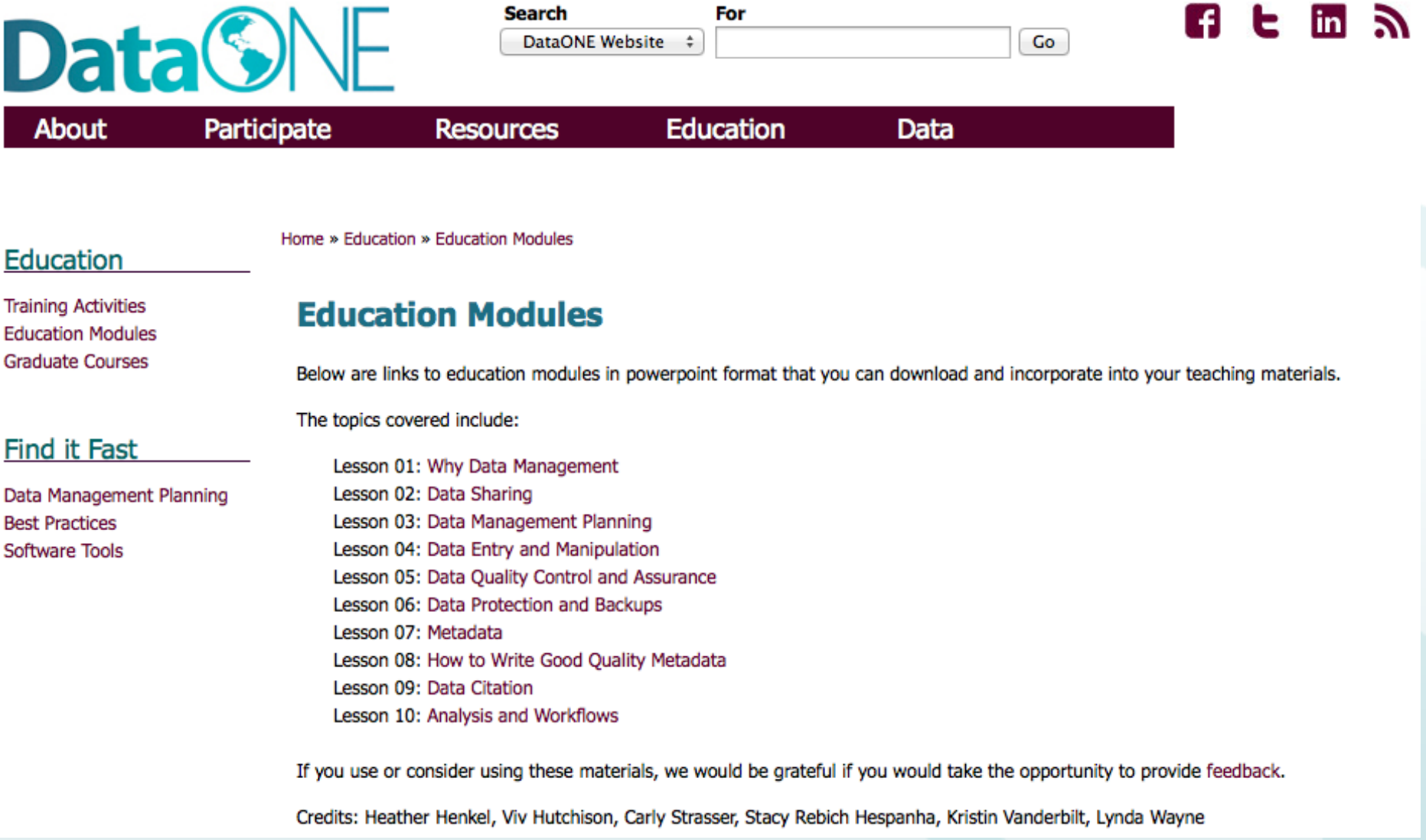
Technical Expertise Required: Basic programming skills
Cost: Cost-basis
Website for more information: Esri - The GIS Software Leader
Additional Information:

- http://www.esri.com/software/arcgis/index.html
- http://en.wikipedia.org/wiki/ArcGIS

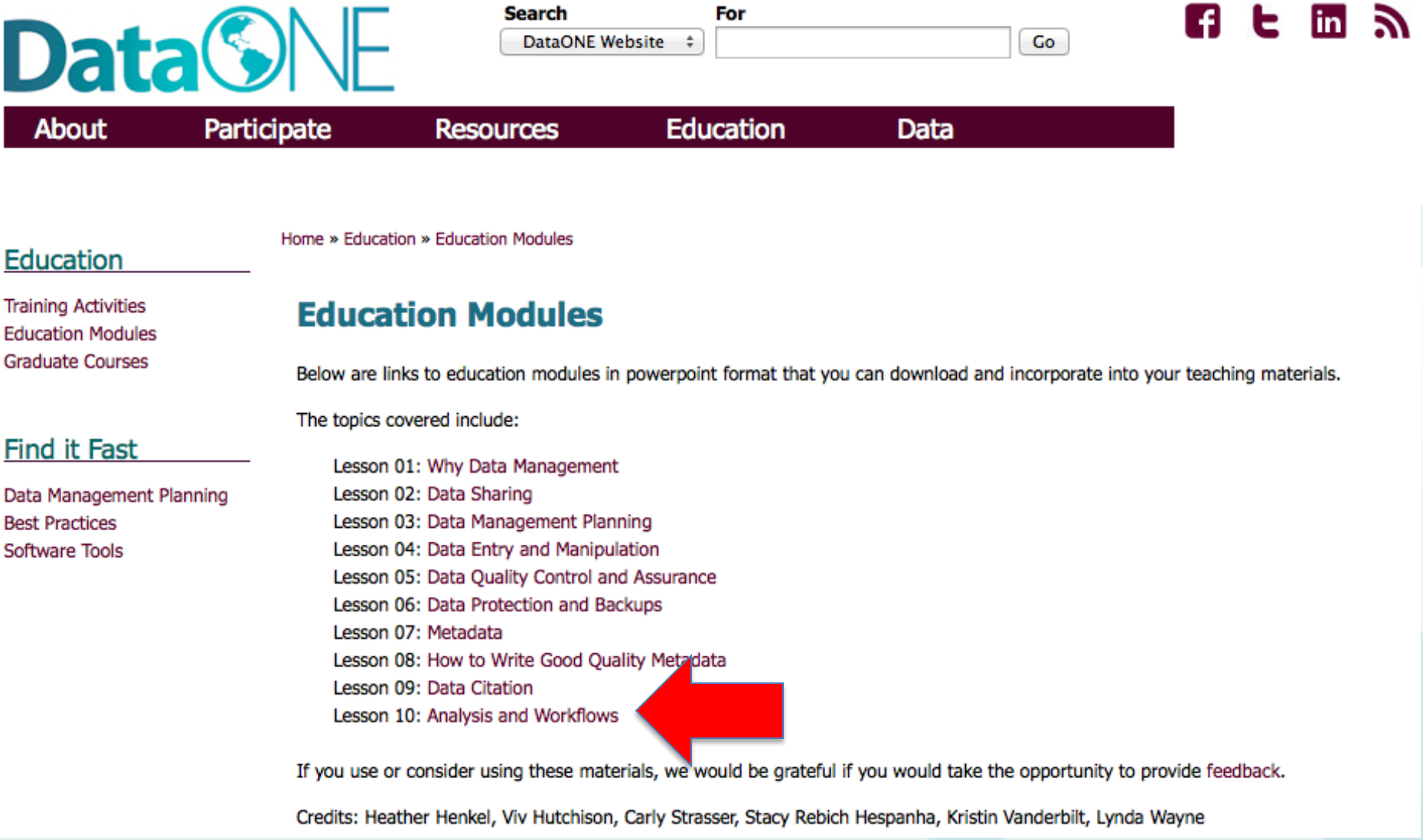
Platform: Windows, UNIX
Tags: GIS map analyze visualization

‹ Previous Next ›

In-depth Training

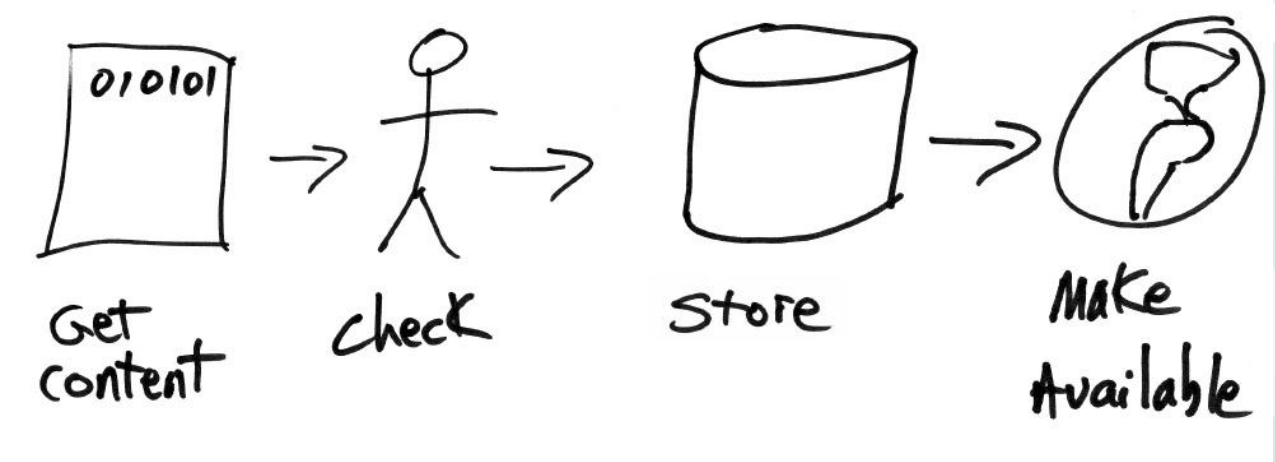


In-depth Training



Tutorials on Data Management

Lesson 10: Analysis and Workflows

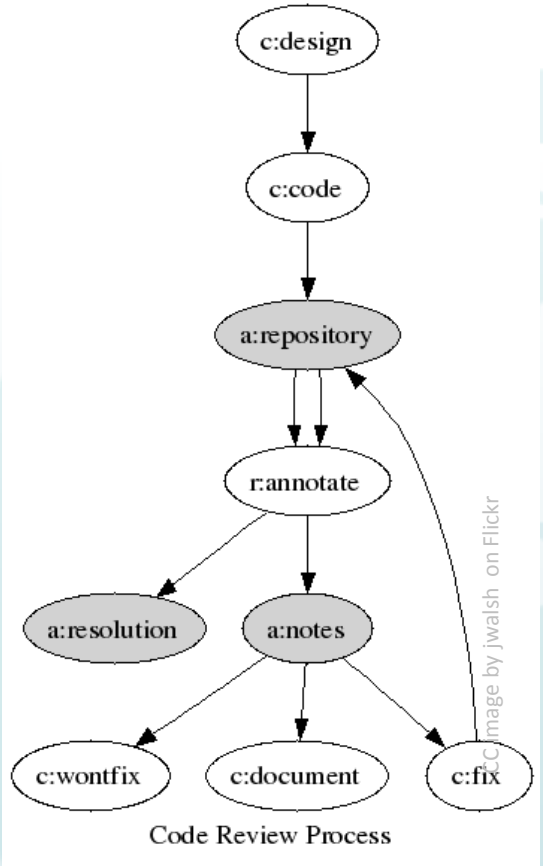


CC image by wlef70 on Flickr

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

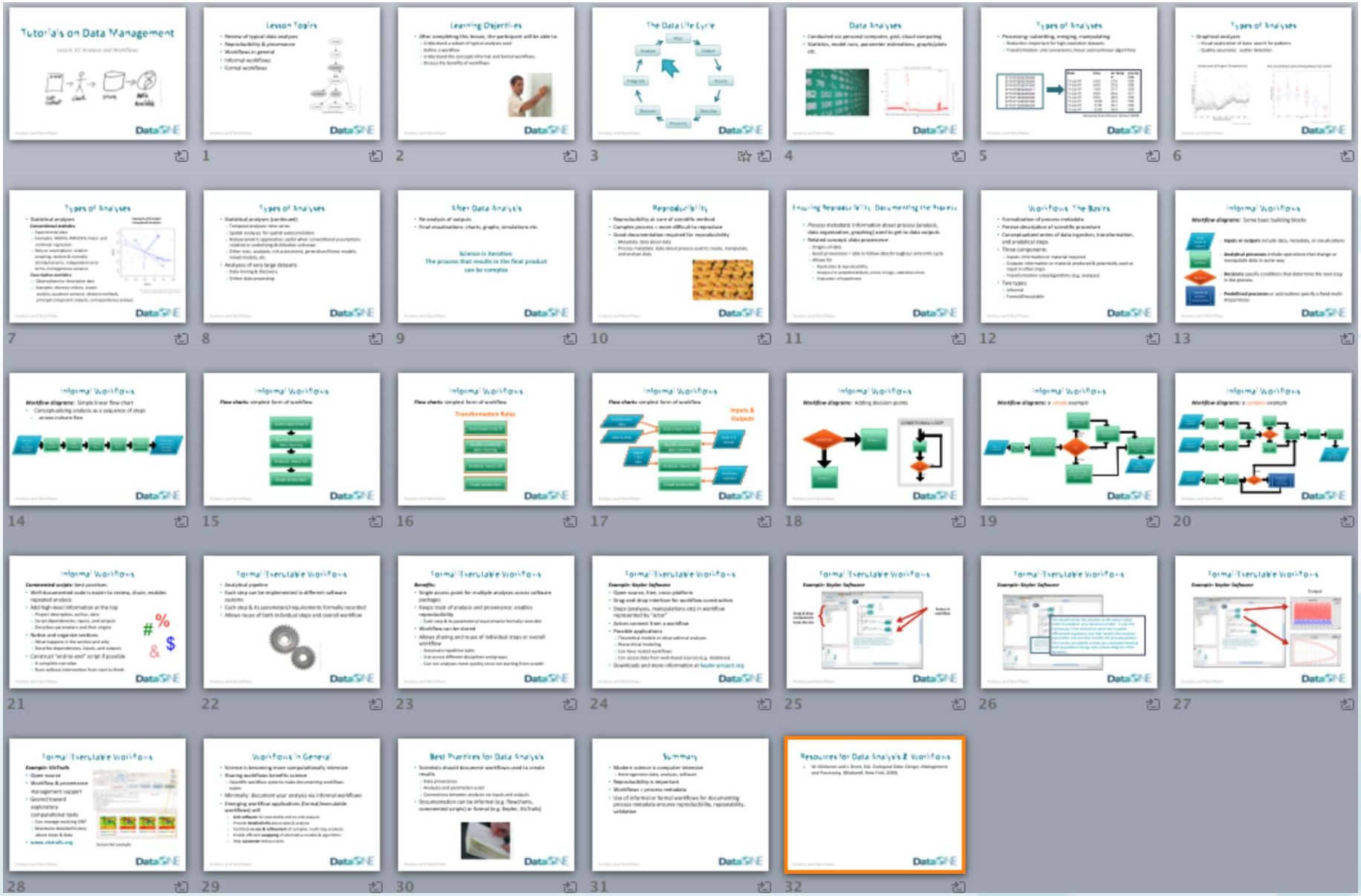


Learning Objectives

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

DataONE

UNM UNIVERSITY LIBRARIES

Made possible by generous support from Dr. Walter E. Dean

Walter E. Dean Environmental Information Management Institute

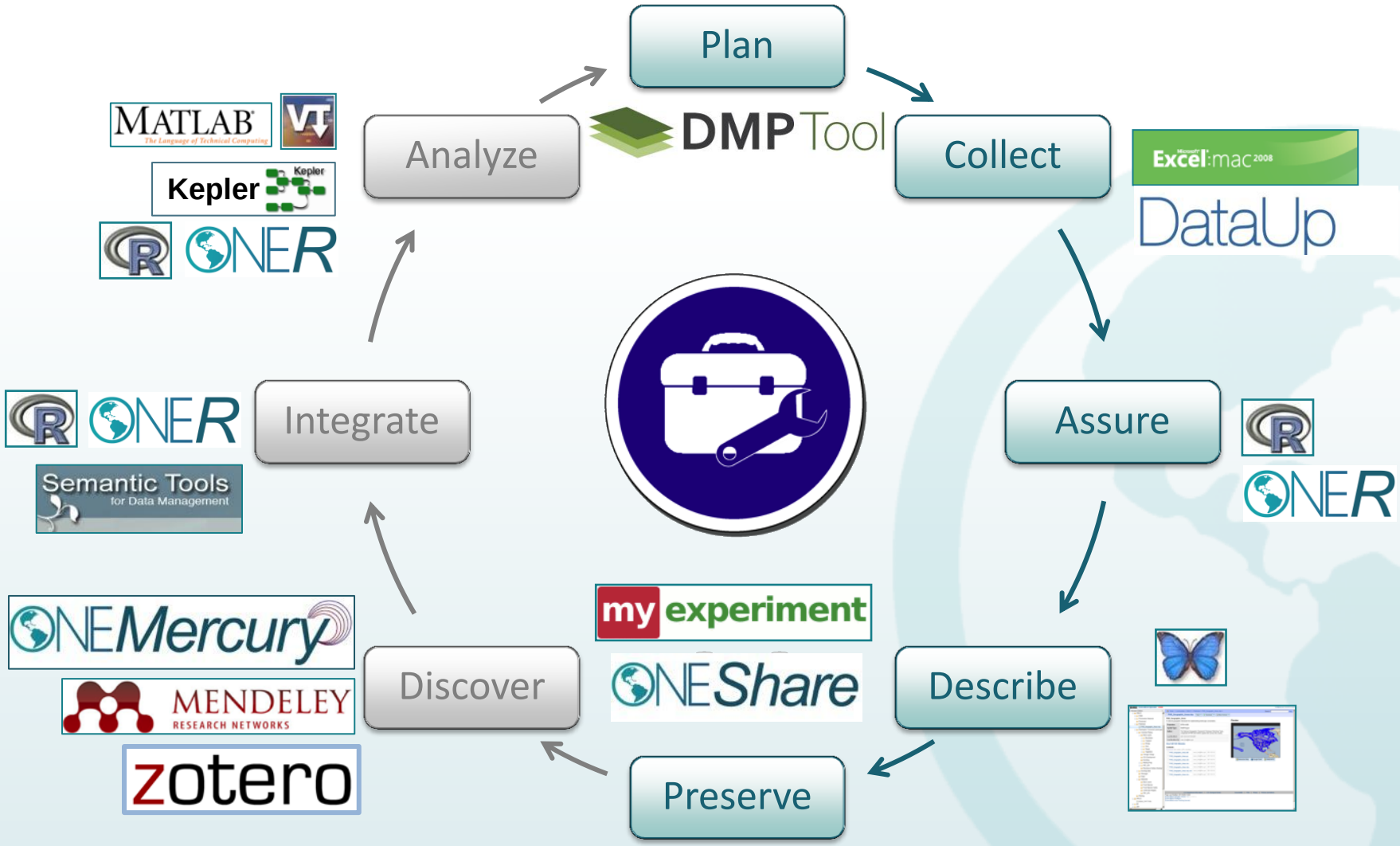
June 3-21, 2013

University of New Mexico

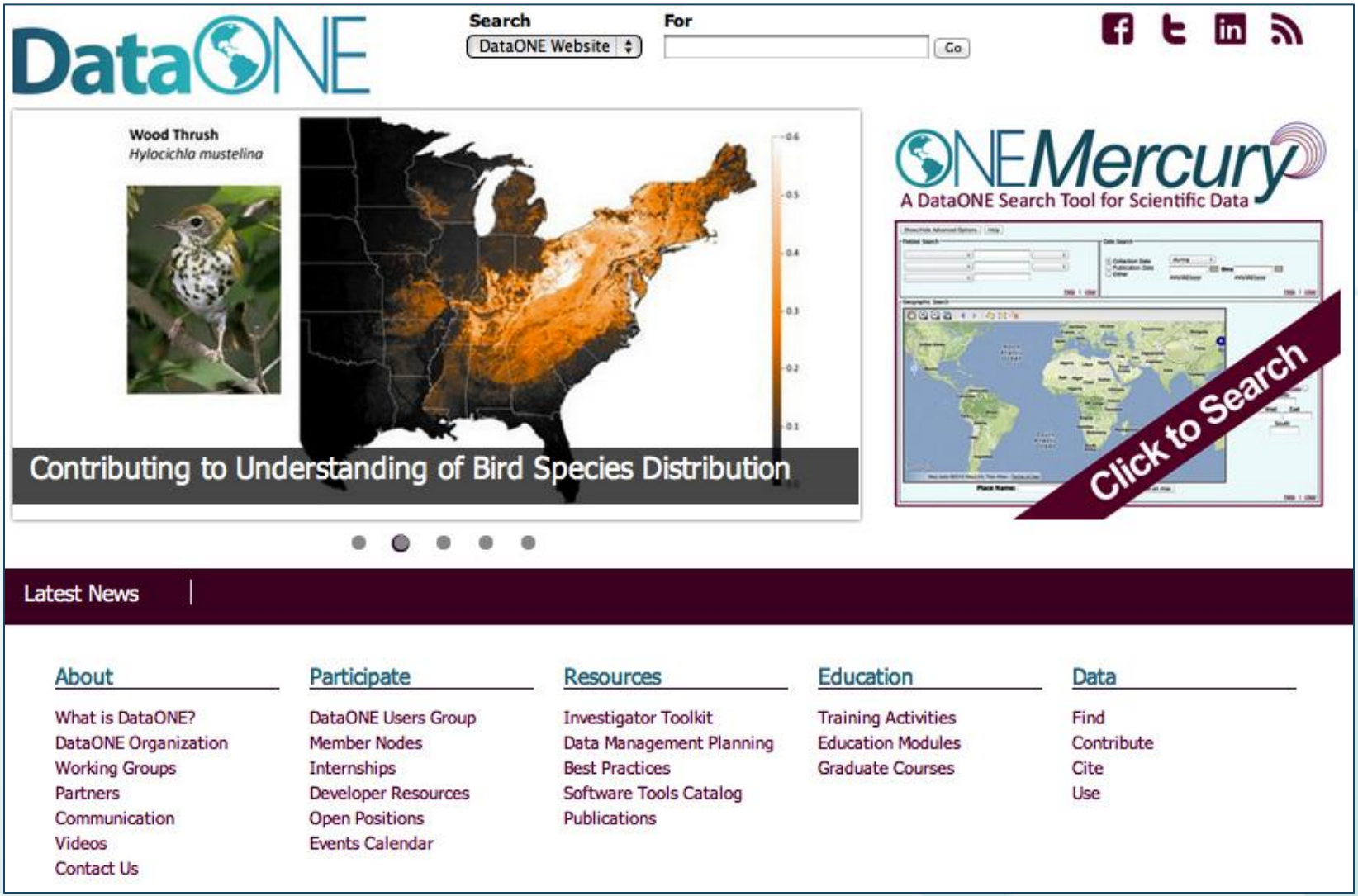
Walter E. Dean Environmental Information Management Institute

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

In-depth Training



DataONE.org


DataONE
Search
DataONE Website
For
Go
Wood Thrush
Hylocichla mustelina
Contributing to Understanding of Bird Species Distribution
ONEMercury
A DataONE Search Tool for Scientific Data
Click to Search
Latest News
About
What is DataONE?
DataONE Organization
Working Groups
Partners
Communication
Videos
Contact Us
Participate
DataONE Users Group
Member Nodes
Internships
Developer Resources
Open Positions
Events Calendar
Resources
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications
Education
Training Activities
Education Modules
Graduate Courses
Data
Find
Contribute
Cite
Use

Credits (Best Practices, Software Tools, Education Modules, EIM Summer Institute)

Best Practices and Software Tools:

Bob Cook, William Michener, Rebecca Koskela, Amber Budden, Carly Strasser, Karl Benedict, Corinna Gries, Christine Laney, Ken Masarie, Mary McCloud, Inigo San Gil, Mark Servilla, Wade Sheldon, Will Shuart, Kristin Vanderbilt, Chris Jones, Cindy Parr, Damien Gessler, Emory Boose, Eric Lind, Faerthen Felix, Jeff Brown, Jeff Horsburgh, Jim Regetz, John Porter, Juliana Freire, Kevin Comerford, Margaret O'Brien, Rebecca Lubas, Robert Olendorf, Robert Stevenson, Ruth Duerr, Steve Tessler, Ted Haberman, Theresa Valentine, Thomas Burley, Trisha Cruse, Todd Grappone, Thorny Staples, Sherry Lake, Sharon Farb, Perry Willett, Michael Grady, Martin Donnelly, Gunter Waibel, Beth Sandore, Andrew Sallans, Marissa Strong, Viv Hutchison

(1) Education Modules and (2) EIM Summer Institute:

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

② Laura Arguelles, Karl Benedict, Robert Cook, Rebecca Koskela, William Michener, Bob Olendorf, John Porter, Jim Regetz, Will Shuart, and Kristin Vanderbilt

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

LEON LEVY FOUNDATION