Curriculum Outline

- Motivation - 1 hour
- DataONE slides and activities - ~6 hours
- Project-specific principles, roles and procedures (TBD) - 4 hours?
Motivation

- Good data management and quality control is a necessary and important precursor for good research results - it is INTEGRAL to productive research
- Example - Whirling Disease research in [this USGS module, starting at slide 12](#)
- Credibility in research results is critical if they are to influence policy sustainably
- [Maximizing Endangered Species Research](#) - outlines some potential impacts of poor research on endangered species advocacy
Motivation

“Several times, I've seen colleagues called to court in order to testify about conditions they have observed.

Without a strong tradition of constant review and approval of basic data, they would've been in deep trouble under cross-examination. Instead, they were able to produce field notes, data approval records, and the like, to back up their testimony.

It's one thing to be questioned by a college student who is working on a project for school. It's another entirely to be grilled by an attorney under oath with the media present.”

- Nelson Williams, Scientist
  US Geological Survey

Why Data Management
Why Data Management

"Climategate"

The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics (in part due to good data management).
Why Manage Data: Researcher Perspective

• Manage your data for yourself:
  o **Keep yourself organized** – be able to find your files (data inputs, analytic scripts, outputs at various stages of the analytic process, etc.)
  o **Track your science processes for reproducibility** – be able to match up your outputs with exact inputs and transformations that produced them
  o **Better control versions of data** – easily identify versions that can be periodically purged
  o **Quality control** your data more efficiently
Why Data Management:
Researcher Perspective

• **To avoid data loss** (e.g. making backups)
• Format your data for **re-use** (by yourself or others)
• **Be prepared**: Document your data for your own recollection, accountability, and re-use (by yourself or others)
• Gain **credibility and recognition** for your science efforts through data sharing!

CC image by UWW ResNet on Flickr
Why Data Management: Foundation to Advance Science

- Data is a valuable asset – it is expensive and time consuming to collect
- Data should be managed to:
  - maximize the effective use and value of data and information assets
  - continually improve the quality including: data accuracy, integrity, integration, timeliness of data capture and presentation, relevance, and usefulness
  - ensure appropriate use of data and information
  - facilitate data sharing
  - ensure sustainability and accessibility in long term for re-use in science
Examples from our Team
Examples from our Team
Examples from our Team
Examples from our Team
Good Data Management Practices for All Team Members

From DataOne Educational Modules

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: Protecting Your Data
*Lesson 07: Metadata
Lesson 08: Data Citation
Lesson 09: Analysis and Workflows
Lesson 10: Legal and Policy Issues
Good Data Management Practices for All Team Members

From DataOne Educational Modules

For each module there are:

- A one page handout (.pdf) - gives a quick look at highlights for the module
- A Powerpoint slide deck (.pptx) - a more detailed look at the subject; many slides contain notes that are useful
- A hands-on exercise (.pdf)
Tutorials on Data Management

Lesson 4: Data Collection, Entry, and Manipulation
Learning Objectives

• Recognize and plan for inconsistencies that can make a dataset difficult to understand and/or manipulate
• Describe characteristics of stable data formats and list reasons for using these formats
• Identify data entry tools
• Identify validation measures that can be performed as data is entered
• Review best practices for data integration
• Describe the basic components of a relational database
Data Entry and Manipulation - Lesson 04

- Slides 1 to 27 recommended for focus (1 hour)
- Activity (1 hour)
- See slides 7, 8, 18, 20 for more detail on what’s covered
Tutorials on Data Management

Lesson 5: Data Quality Control and Assurance
Learning Objectives

• After completing this lesson, the participant will be able to:
  o Define data quality control and data quality assurance
  o Perform quality control and assurance on their data at all stages of the research cycle
Data Quality Control and Assurance - Lesson 05

- All slides recommended (1 hour)
- Activity (45 minutes to 1 hour)
- See slides 5, 7, 11 for more detail on what's covered
What is Metadata

Lesson 7: Metadata
Learning Objectives

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

• Identify and list the types of information typically included in metadata records for environmental datasets
• Identify 3 reasons metadata is of value to data users, data developers, and organizations
• List 3 uses for metadata, beyond discovery of data
• Identify and describe factors that may determine which metadata standards are most appropriate for a given dataset
• List steps to prepare to write metadata
• Explain how to write good metadata
Metadata - Lesson 07

- All slides recommended; can skip 14-16, move quickly through 26-30 and 36-42 (1 hour)
- Activity (45 minutes to 1 hour)
- See slides 5,17,18,23,45 for more detail on what’s covered
Last Thoughts on DataONE modules

- Though trying to focus this curriculum on specified pain points, all of these modules could be useful!
- Lesson 6 on ‘Protecting Your Data’ could be particularly useful for those responsible for mitigating the risk of data loss (e.g. back-ups)
# Databases for Many Majors

An interactive learning module for (easily!) teaching ecology students (and professors) about databases for managing and querying large datasets

**Becky A. Ball**, **Suzanne W. Dietrich**, and **Don Goelman**

1School of Mathematical and Natural Sciences, Arizona State University
2Department of Computing Sciences, Villanova University

---

## Purpose

To introduce database skills to ecologists so that they can more efficiently handle large datasets.

- Working with large datasets is an increasingly integral part of ecological research. Students need to be trained in tools & techniques to handle and analyze large amounts of data.
- Most ecologists probably use spreadsheets (e.g. Excel), which is not the same as a proper database. They are clumsy and inefficient!
- True databases are more efficient, easier to manage, and provide a powerful tool to ask different questions, or "queries", of that one dataset without constantly changing the spreadsheet (e.g., without re-sorting, copy & pasting, repeatedly typing equations, etc.).
- But it is difficult to teach ecology students about databases, when likely their professors are not experts in databases either.

## Using the Introductory Modules

The modules can be found online at: [http://databasesmanymajors.faculty.asu.edu/](http://databasesmanymajors.faculty.asu.edu/)

The tabs at the top of the page represent the interactive modules that engage student learning in three major topics:

1) **Introduction to Databases**: Promotes the understanding of how a database operates, and how to better utilize the power databases have. Introduces the limitations of spreadsheets in dealing with redundancies, anomalies, missing data, etc.

2) **Introduction to Querying**: Provides a conceptual introduction to the various operations required to retrieve data from a database to answer a question, and introduces the SQL language to formulate these operations.

3) **Conceptual Database Design**: Builds on the previous two modules with the introduction to the conceptual design of data, giving an overview to mapping relationships among data entities to design a database based on your dataset.

For each module, you can view either the stock version, or scroll down to the "Customizations" section to select the...

## Why databases?

- Spreadsheets (e.g. Excel) are clunky and inefficient compared to a proper database (e.g. Access).
- A database breaks the dataset down into tables, which are more easily managed.
- The content of the tables are linked, according to your design.
Data Management Principles for this Team

1. Treat data securely, as they are about endangered species.

2. Consistency is essential. Everyone handling data must use the same abbreviations, spelling, format, and store data in the same databases. This consistency includes file naming and directory hierarchies.

3. Every step in the data handling and manipulation process should be documented. This documentation (metadata) will help us answer our research questions in the long run and is also required by our USGS collaborators.

4. Active project data management should be conducted ‘little and often,’ and not left to pile up. For example, data should be entered into electronic form promptly.

5. If systems or procedures are hindering successful data management and data quality control, it is your responsibility to inform the Field Crew Team Leader (and Project Leader as appropriate).
Data Management Roles for this Team

● Every field worker is responsible for recording complete, accurate, and legible field notes/data sheets.

● Every field worker is responsible for accurately transcribing data into electronic data files.

● Everyone transcribing data is responsible for ensuring that the file contains metadata that defines all abbreviations, describes all column headings.
Data Management Roles for this Team

- Wildlife Technician 1 is responsible for maintaining bog frog data files, Wildlife Technician 2 is responsible for maintaining gopher tortoise and red-cockaded woodpecker data files, and Field crew leader is responsible for maintaining flatwoods salamander data files.

- Field crew leader is responsible for ensuring that all data are proofed, conform to project standards, contain complete metadata, are labeled correctly, are stored correctly, and are backed up.

- Field crew leader and postdoc are responsible for ensuring that complete metadata conform to USGS standards and that as well as archiving our complete data file, we create a data file for public distribution with location data redacted.

- [Roles of local research colleagues]
Data Management Procedures

We want the field team member’s input on how best to improve data management and quality...

- What issues are most hampering the field team’s efforts in entering and aggregating data gathered in the field?
- How best to assure data is entered into the master Access database, and done so promptly, consistently and correctly?