Eglin AFB Field Projects Data Management Training Curriculum

C. Haas, G. Brooks, Fish and Wildlife Conservation
J. Petters, University Libraries
J. Smith, Biological Sciences

September 15 2017

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

“Climategate”

guardian.co.uk

Search | Environment | Search

News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment | TV | Video | Community | Blogs | Jobs

Environment › Hacked climate science emails

Climategate scientists cleared of manipulating data on global warming

Muir Russell report says scientists did not fudge data, but they should have been more open about their work

- Read the full text of the review here
- 'Climategate' report - main findings

(86)

Tweet this (99)

David Adam, environment correspondent
The Guardian, Thursday 8 July 2010
Article history

Muir Russell during the release of his report into the scandal of the hacked emails sent by climate scientists from University of East Anglia. Photograph: Sang Tan/AP

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, but a review found they had failed to be open enough about their work.

Sir Muir Russell, the senior civil servant who led a six-month inquiry into the affair, said the "rigour and honesty" of the scientists at the Climatic Research Unit (CRU) at the University of East Anglia (UEA) were not in doubt. His investigation concluded they did not subvert the peer review process to censor criticism and that key data was freely available and could be used by any "competent" researcher.

Environm
Hacked cl
emails · C
Climate ch
scepticism

Science
Climate ch

Educatio
University
Higher ed

UK news

Technology
Hacking

More news

More on this story

UEA's delayed response to climate emails caused by shock, says professor
Former head of research unit responds to criticism by arguing for necessity of assessing excerpts by

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*)

On Environment

Most viewed | Zeitgeist | Latest

Last 24 hours

1. BP oil spill mostly cleaned up, says US
2. Battle to halt BP oil spill is nearing its end, says Barack Obama

Why Manage Data: Researcher Perspective

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

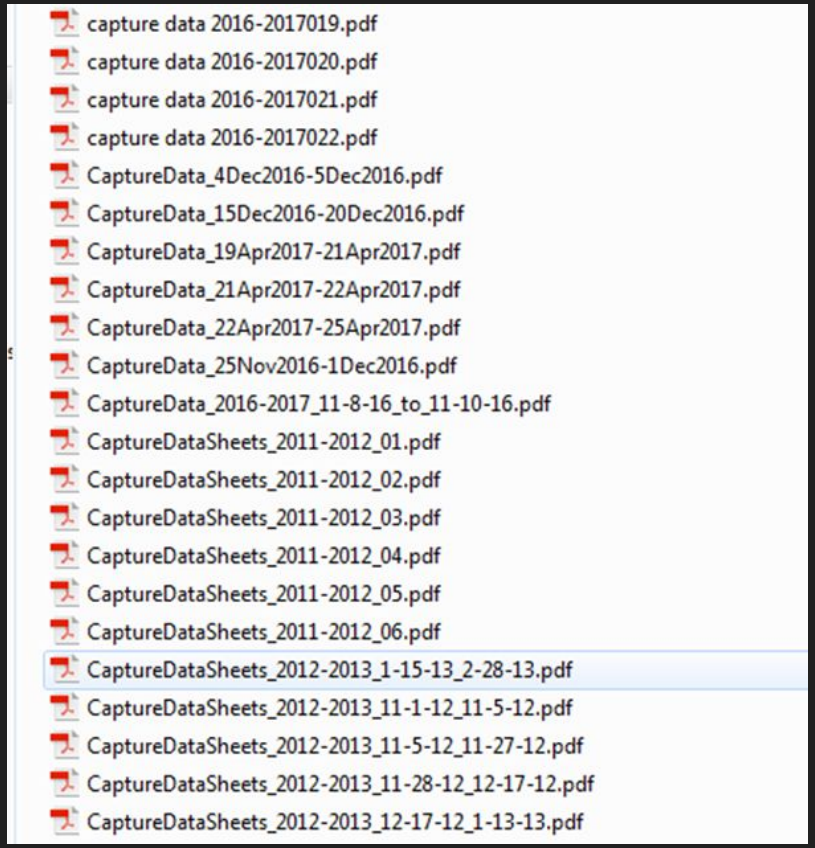
Why Data Management: Researcher Perspective

- **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!

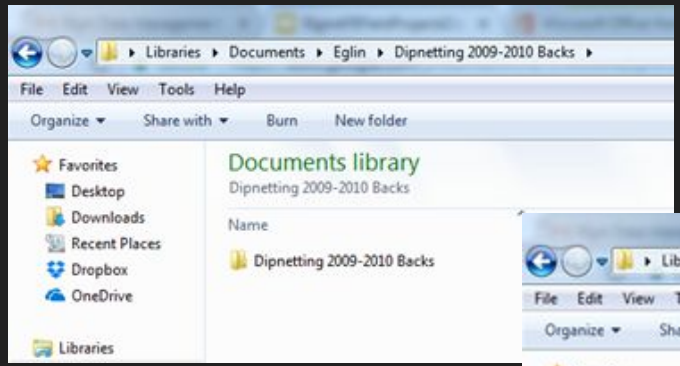
Why Data Management: Foundation to Advance Science

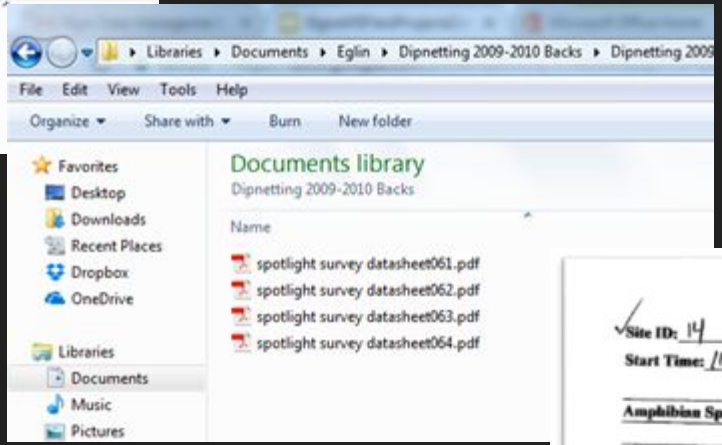
- Data is a valuable asset – it is expensive and time consuming to collect
- Data should be managed to:
  - **maximize the effective use and value** of data and information assets
  - continually **improve the quality** including: data accuracy, integrity, integration, timeliness of data capture and 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

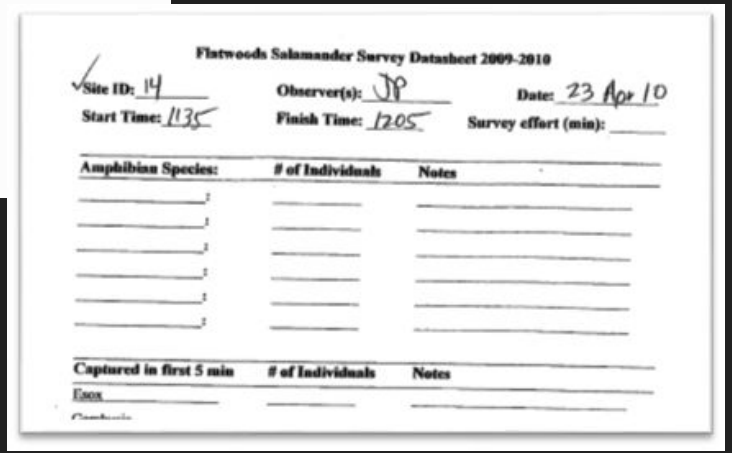
Examples from our Team



Examples from our Team







Flatwoods Salamander Survey Datasheet 2009-2010

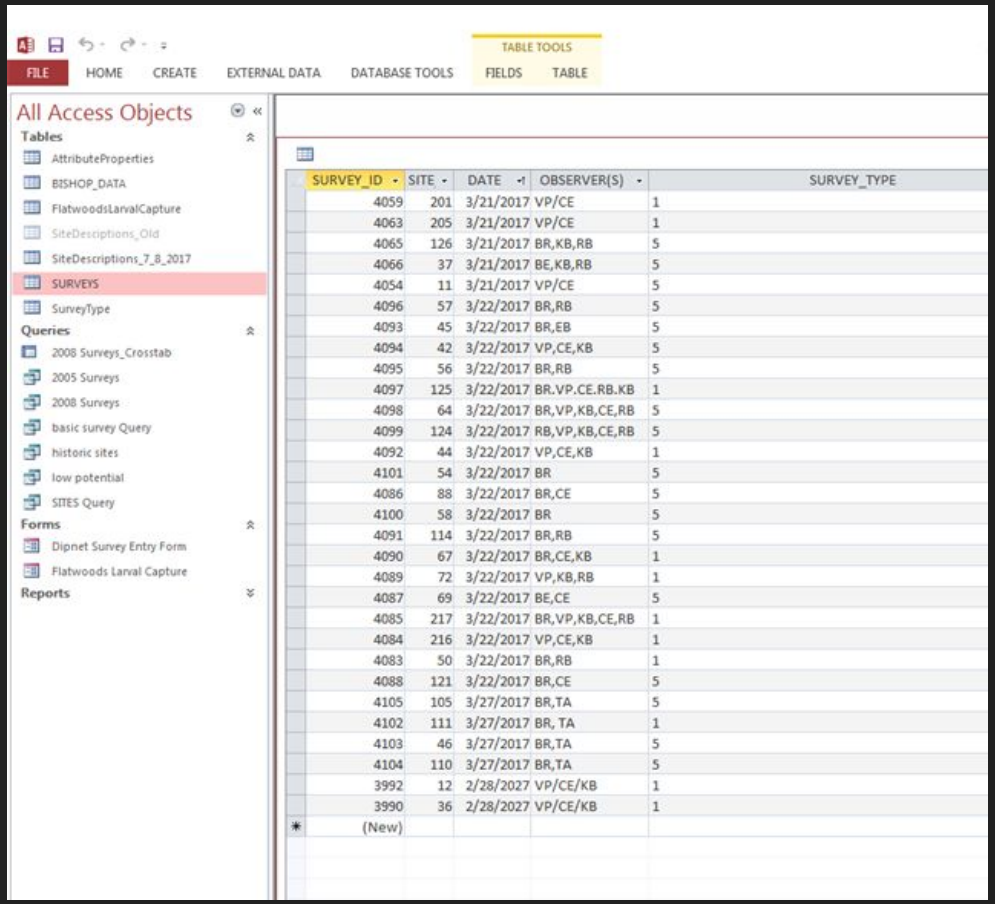
Site ID: 14 Observer(s): JP Date: 23 Apr 10

Start Time: 1135 Finish Time: 1205 Survey effort (min): ____

Amphibian Species:	# of Individuals	Notes

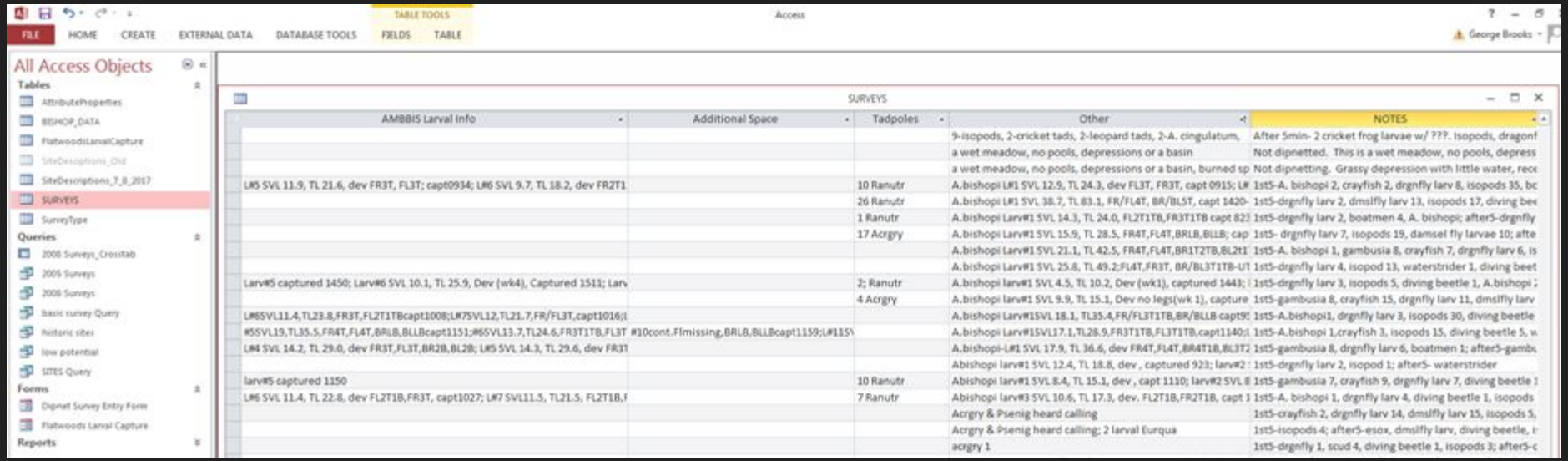
Captured in first 5 min	# of Individuals	Notes
Esox		

Examples from our Team



SURVEY_ID	SITE	DATE	OBSERVER(S)	SURVEY_TYPE
4059	201	3/21/2017	VP/CE	1
4063	205	3/21/2017	VP/CE	1
4065	126	3/21/2017	BR,KB,RB	5
4066	37	3/21/2017	BE,KB,RB	5
4054	11	3/21/2017	VP/CE	5
4096	57	3/22/2017	BR,RB	5
4093	45	3/22/2017	BR,EB	5
4094	42	3/22/2017	VP,CE,KB	5
4095	56	3/22/2017	BR,RB	5
4097	125	3/22/2017	BR.VP.CE.RB.KB	1
4098	64	3/22/2017	BR,VP,KB,CE,RB	5
4099	124	3/22/2017	RB,VP,KB,CE,RB	5
4092	44	3/22/2017	VP,CE,KB	1
4101	54	3/22/2017	BR	5
4086	88	3/22/2017	BR,CE	5
4100	58	3/22/2017	BR	5
4091	114	3/22/2017	BR,RB	5
4090	67	3/22/2017	BR,CE,KB	1
4089	72	3/22/2017	VP,KB,RB	1
4087	69	3/22/2017	BE,CE	5
4085	217	3/22/2017	BR,VP,KB,CE,RB	1
4084	216	3/22/2017	VP,CE,KB	1
4083	50	3/22/2017	BR,RB	1
4088	121	3/22/2017	BR,CE	5
4105	105	3/27/2017	BR,TA	5
4102	111	3/27/2017	BR, TA	1
4103	46	3/27/2017	BR,TA	5
4104	110	3/27/2017	BR,TA	5
3992	12	2/28/2027	VP/CE/KB	1
3990	36	2/28/2027	VP/CE/KB	1
(New)				

Examples from our Team



AMBBIS Larval Info	Additional Space	Tadpoles	Other	NOTES
			9-isopods, 2-cricket tads, 2-leopard tads, 2-A. cingulatum,	After 5min- 2 cricket frog larvae w/ ???. Isopods, dragonf
			a wet meadow, no pools, depressions or a basin	Not dipnetted. This is a wet meadow, no pools, depress
			a wet meadow, no pools, depressions or a basin, burned sp	Not dipnetting. Grassy depression with little water, reco
L#5 SVL 11.9, TL 21.6, dev FR3T, FL3T; capt0934; L#6 SVL 9.7, TL 18.2, dev FR2T1		10 Ranutr	A.bishopi L#1 SVL 12.9, TL 24.3, dev FL3T, FR3T, capt 0915; L#	1st5-A. bishopi 2, crayfish 2, drgnfly larv 8, isopods 35, bc
		26 Ranutr	A.bishopi L#1 SVL 38.7, TL 83.1, FR/FL4T, BR/BL5T, capt 1420-	1st5-drgnfly larv 2, dmslfly larv 13, isopods 17, diving bee
		1 Ranutr	A.bishopi Larv#1 SVL 14.3, TL 24.0, FL2T1TB,FR3T1TB capt 823	1st5-drgnfly larv 2, boatmen 4, A. bishopi; after5-drgnfly
		17 Acrgry	A.bishopi Larv#1 SVL 15.9, TL 28.5, FR4T,FL4T,BRLB,BLLB; cap	1st5- drgnfly larv 7, isopods 19, damsel fly larvae 10; afte
			A.bishopi Larv#1 SVL 21.1, TL 42.5, FR4T,FL4T,BR1T2TB,BL2T1	1st5-A. bishopi 1, gambusia 8, crayfish 7, drgnfly larv 6, is
			A.bishopi Larv#1 SVL 25.8, TL 49.2;FL4T,FR3T, BR/BL3T1TB-U1	1st5-drgnfly larv 4, isopod 13, waterstrider 1, diving beet
Larv#5 captured 1450; Larv#6 SVL 10.1, TL 25.9, Dev (wk4), Captured 1511; Larv		2; Ranutr	A.bishopi larv#1 SVL 4.5, TL 10.2, Dev (wk1), captured 1443;	1st5-drgnfly larv 3, isopods 5, diving beetle 1, A.bishopi 2
		4 Acrgry	A.bishopi larv#1 SVL 9.9, TL 15.1, Dev no legs(wk 1), capture	1st5-gambusia 8, crayfish 15, drgnfly larv 11, dmslfly larv
L#6SVL11.4,TL23.8,FR3T,FL2T1TBcapt1008;L#7SVL12,TL21.7,FR/FL3T,capt1016;L			A.bishopi Larv#1SVL 18.1, TL35.4,FR/FL3T1TB,BR/BLLB capt9	1st5-A.bishopi1, drgnfly larv 3, isopods 30, diving beetle
#5SVL19,TL35.5,FR4T,FL4T,BRLB,BLLBcapt1151;#6SVL13.7,TL24.6,FR3T1TB,FL3T	#10cont.Flmissing,BRLB,BLLBcapt1159;L#11S		A.bishopi Larv#1SVL17.1,TL28.9,FR3T1TB,FL3T1TB,capt1140;1	1st5-A.bishopi 1,crayfish 3, isopods 15, diving beetle 5, w
L#4 SVL 14.2, TL 29.0, dev FR3T,FL3T,BR2B,BL2B; L#5 SVL 14.3, TL 29.6, dev FR3T			A.bishopi-L#1 SVL 17.9, TL 36.6, dev FR4T,FL4T,BR4T1B,BL3T2	1st5-gambusia 8, drgnfly larv 6, boatmen 1; after5-gambu
			Abishopi larv#1 SVL 12.4, TL 18.8, dev , captured 923; larv#2	1st5-drgnfly larv 2, isopod 1; after5- waterstrider
larv#5 captured 1150		10 Ranutr	Abishopi larv#1 SVL 8.4, TL 15.1, dev , capt 1110; larv#2 SVL 8	1st5-gambusia 7, crayfish 9, drgnfly larv 7, diving beetle 1
L#6 SVL 11.4, TL 22.8, dev FL2T1B,FR3T, capt1027; L#7 SVL11.5, TL21.5, FL2T1B,F		7 Ranutr	Abishopi larv#3 SVL 10.6, TL 17.3, dev. FL2T1B,FR2T1B, capt 1	1st5-A. bishopi 1, drgnfly larv 4, diving beetle 1, isopods
			Acrgry & Psenig heard calling	1st5-crayfish 2, drgnfly larv 14, dmslfly larv 15, isopods 5,
			Acrgry & Psenig heard calling; 2 larval Eurqua	1st5-isopods 4; after5-esox, dmslfly larv, diving beetle,
			acrgry 1	1st5-drgnfly 1, scud 4, diving beetle 1, isopods 3; after5-c

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

DataONE

Learning Objectives

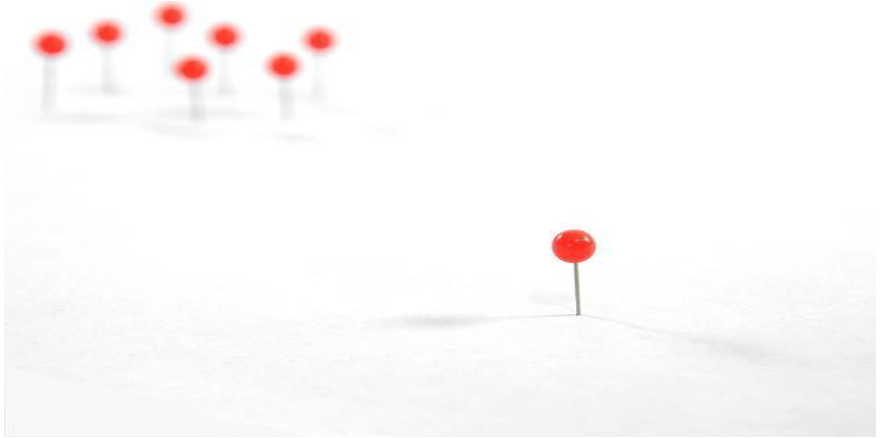
- Recognize and plan for inconsistencies that can make a dataset difficult to understand and/or manipulate
- Describe characteristics of stable data formats and list reasons for using these formats
- Identify data entry tools
- Identify validation measures that can be performed as data is entered
- Review 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

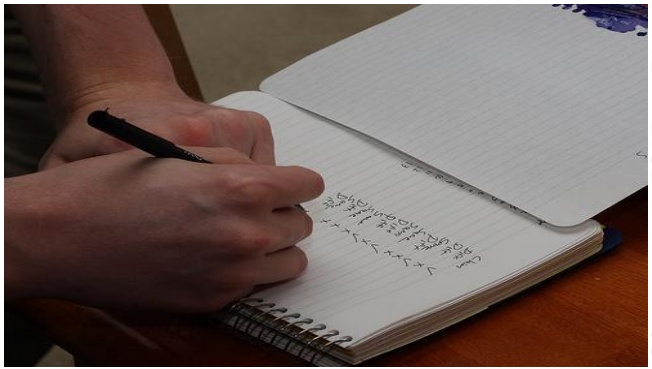
Lesson 5: Data Quality Control and Assurance



DataONE

Learning Objectives

- After completing this lesson, the participant will be able to:
  - Define data quality control and data quality assurance
  - 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

Tutorials on Data Management

Lesson 7: Metadata

What is Metadata

DataONE

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[1], Suzanne W. Dietrich[1], and Don Goelman[2]

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

ASU School of Mathematical and Natural Sciences Arizona State 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 clunky 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/

The tabs at the top of the page represent the interactive modules that engage student learning in the 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.

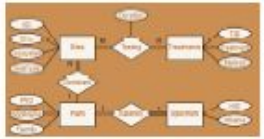
3) **Conceptual Database Design:** Builds on the previous two modules with the introduction to the conceptual design of data, giving an overview to mapping relationships among data entities to design a database based on your dataset.

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?