Research Data Management in Environmental Science

Ali Krzton, Research Data Management Librarian
RBD Library

Overview

- Research data management is at the heart of the scientific method
- It has become an institutional priority
- Environmental sciences are particularly data intensive
- Changing research practices require new data management strategies
- Upfront investment in good systems has huge payoffs later
- Our valuable environmental data is worth it

But first, some history

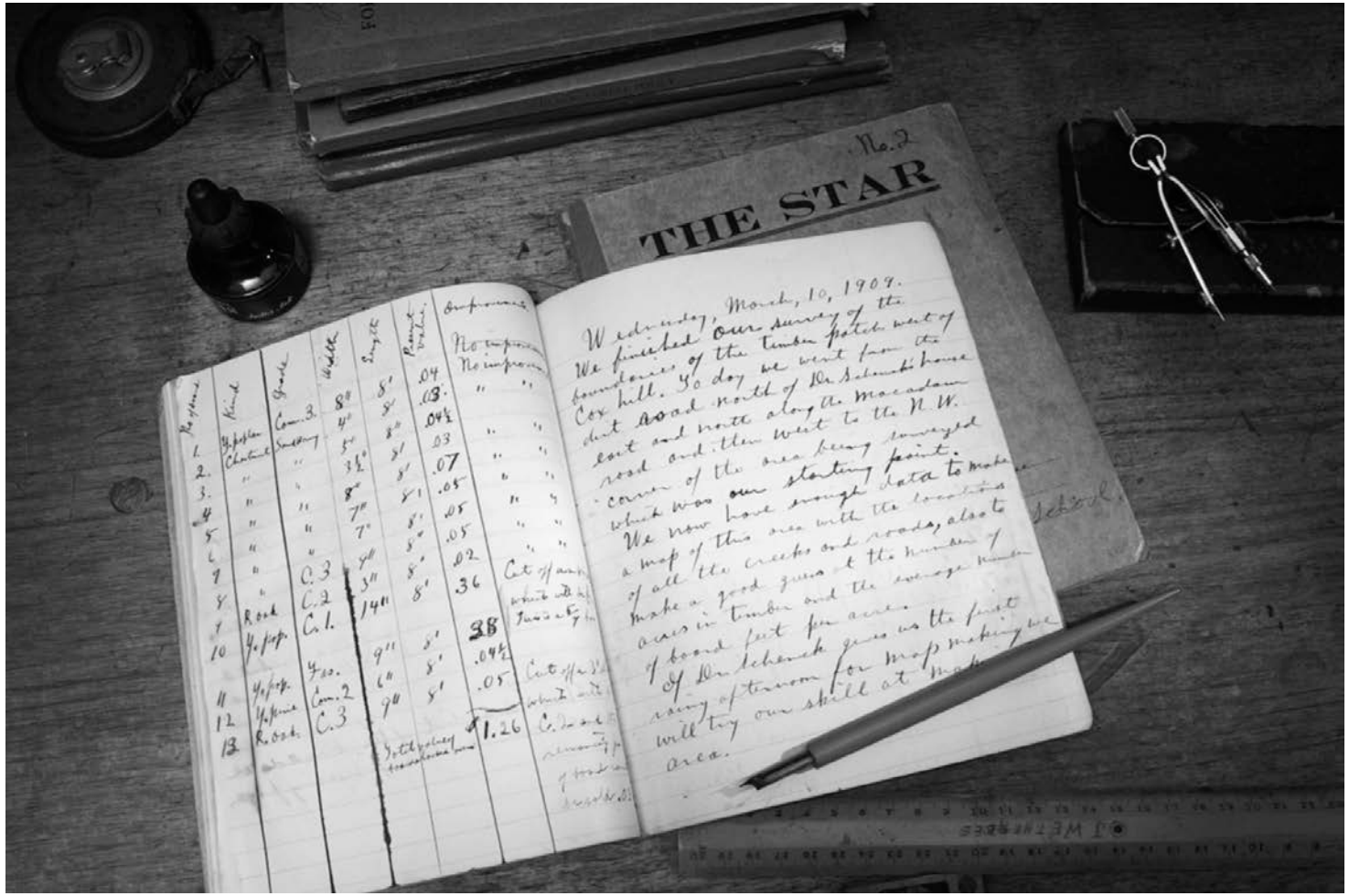


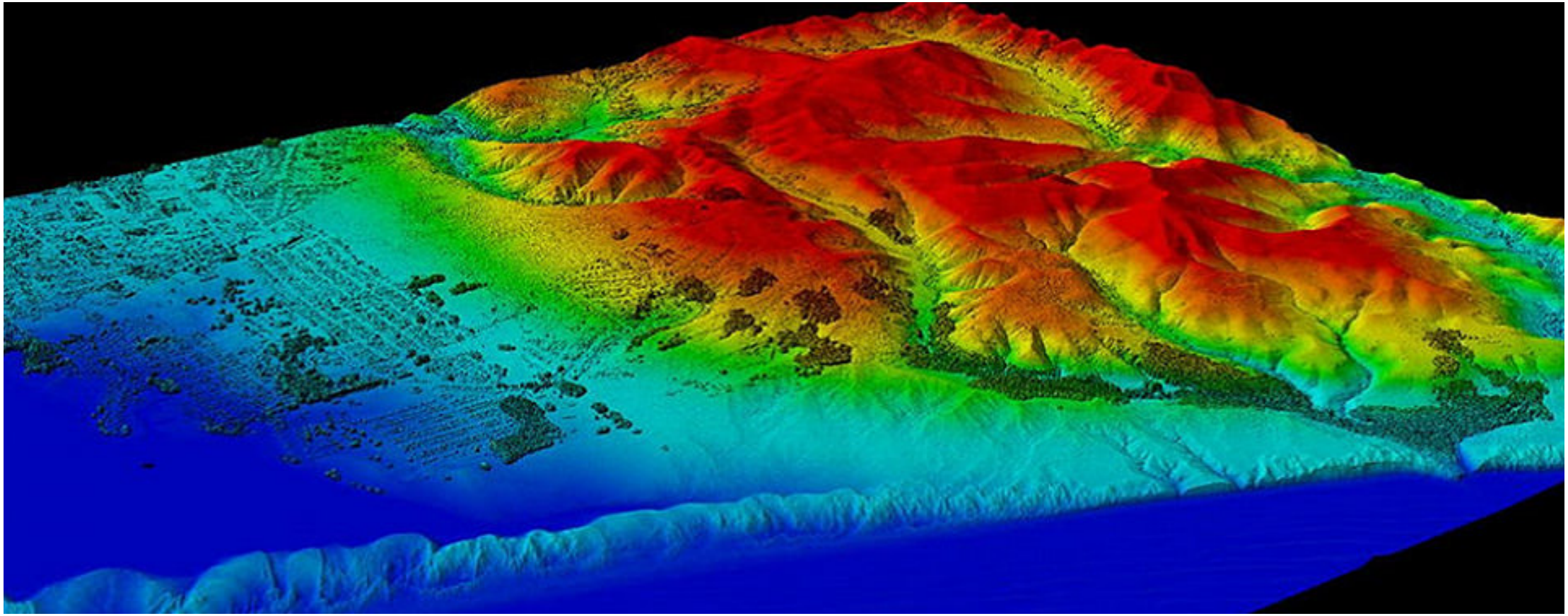
Figure reproduced from Lee (2012)

Features of environmental science

- Research conducted at a wide variety of scales – spatial, temporal, micro to macro levels of analysis
- Aggregated datasets with multiple inputs are the norm
- Data often cannot be recreated (e.g. historical data)
- Decision-makers rely upon our data

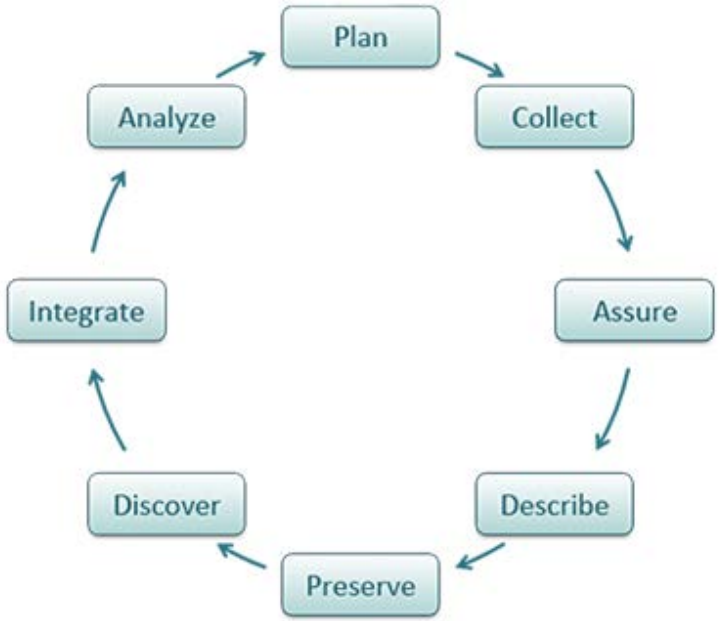

ALL-WEATHER
FIELD BOOK

Our work is becoming more computational



Why research data management now?

- Funding mandates (NIH, NSF, OSTP)
- Journal mandates
- Reproducibility crisis
- Collaborative research models



A sample of RDM issues

- Formatting
- QA/QC
- Software interoperability
- Documentation
- Loss/corruption

Ecology and Evolution
Open Access

Volume 7, Issue 17
September 2017
Pages 6863-6870

ORIGINAL RESEARCH | Open Access

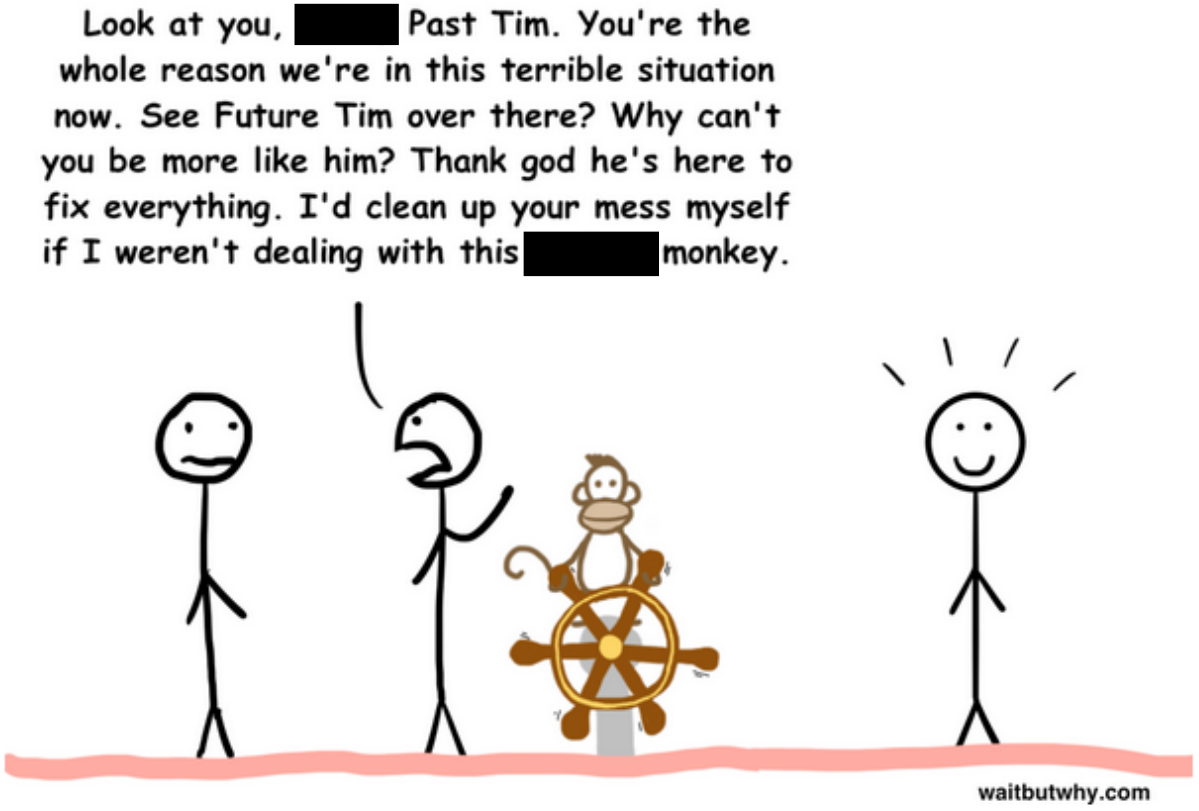
Temporal degradation of data limits biodiversity research

Geiziane Tessarolo, Richard Ladle, Thiago Rangel, Joaquin Hortal

First published: 27 July 2017 | https://doi.org/10.1002/ece3.3259 | Cited by:2

Abstract

Spatial and/or temporal biases in biodiversity data can directly influence the utility, comparability, and reliability of ecological and evolutionary studies. While the effects of biased spatial coverage of biodiversity data are relatively well known, temporal variation in data quality (i.e., the congruence between recorded and actual information) has received much less attention. Here, we develop a conceptual framework for understanding the influence of time on biodiversity data quality based on three main processes: (1) the natural dynamics of ecological systems—such as species turnover or local extinction; (2) periodic taxonomic revisions, and; (3) the loss of physical and



The biggest beneficiary of good research data management...is future you.

Solutions at the micro level

- Consistency within and between
- Talk about research data management as part of scientific procedure
- Set policies within the working group – get buy-in!
- Match training to responsibilities
- Make sure digitized data is machine-readable

What’s wrong with this picture?

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	riuSlado	"	1+	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

Poor data entry – common errors

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

Inconsistencies abound! Location and format of date information, column headings, and column order are different, as are site spellings and codes used. Information has also jumped into other columns where it doesn't belong.

Improved data organization

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.92	got away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping Sheet3

General rules for tabular data

- 1. Each variable must have its own column.
- 2. Each observation must have its own row.
- 3. Each value must have its own cell.

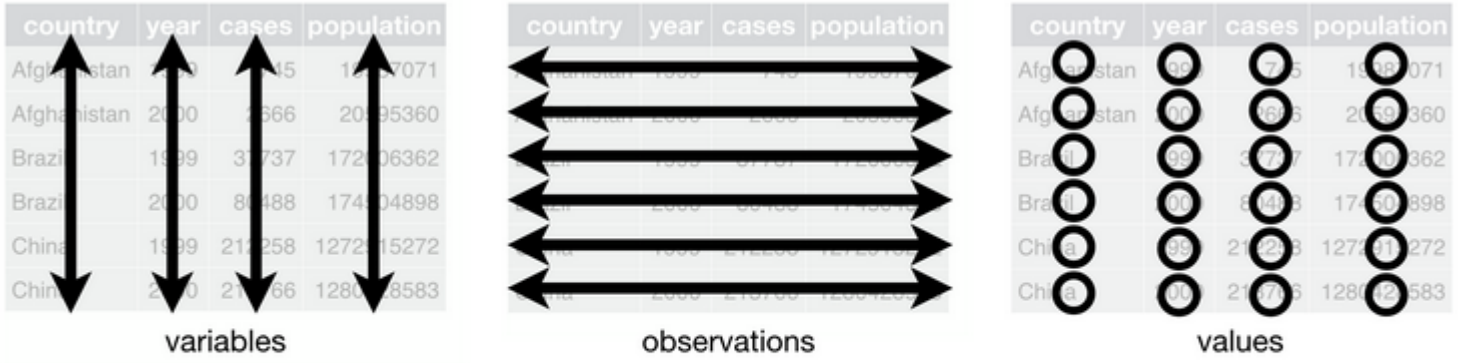


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

File formats

- Some file formats will be determined by equipment or software
- Proprietary file formats can cause problems
- Choose the most compatible formats when possible
- Consider saving a backup/archival copy in an open format
- READMEs should be in plain text (.txt)

File naming – best practices

- Naming needs to be descriptive and consistent
- A unique identifier system may help, but check with collaborators
- Dates and version numbers could also be included
- Shorter is usually better, and _underscores_ or CamelCase may improve readability
- Make sure not to use special characters: offenders include hyphens, periods other than the one before the extension, spaces
- Avoid generic names and words like “new” or “final”

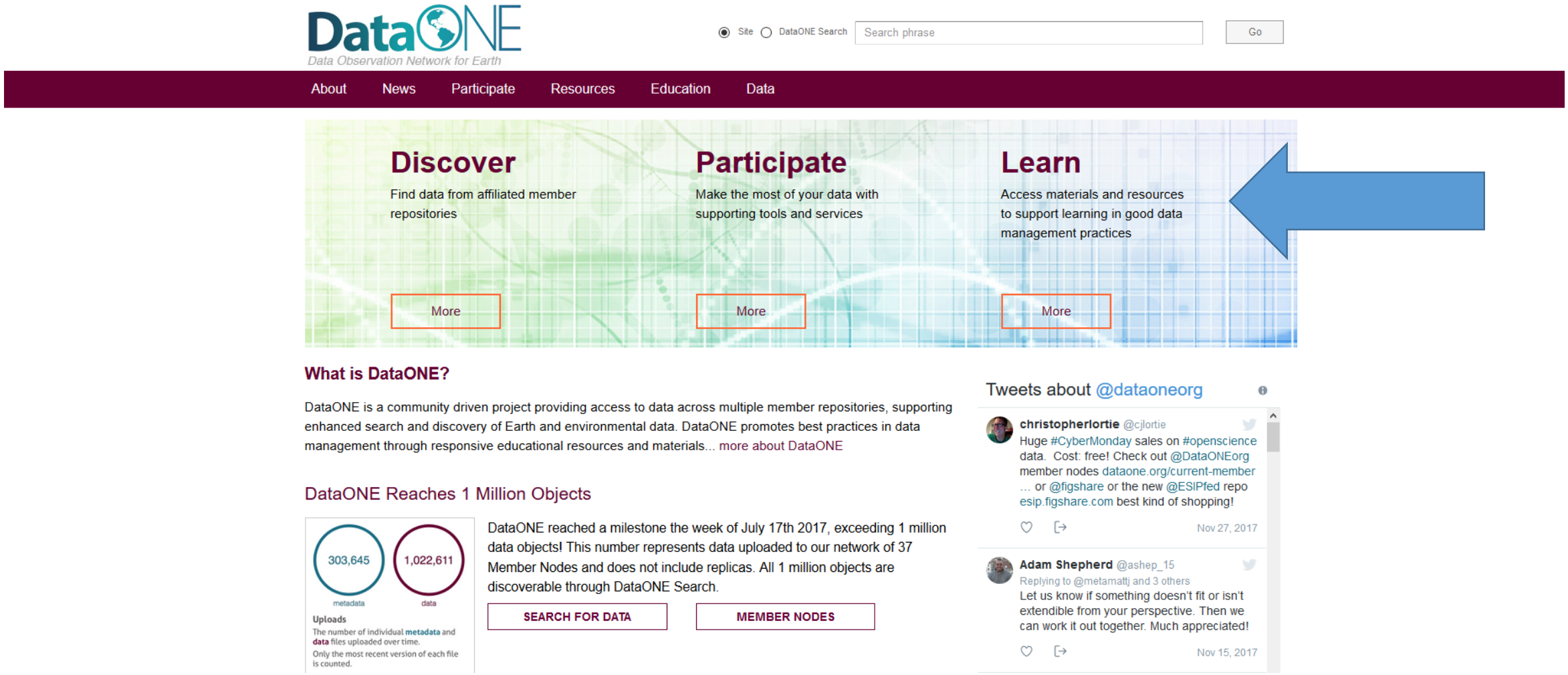
Backups and versioning

- Keeping multiple copies in multiple locations, if possible
- Backup and versioning processes are best automated
- Consider cloud services and/or central repositories
- Version retention: value vs. confusion
- Tools and protocols change too!

Solutions at the macro level

- Working groups and institutions have developed tools designed to improve data management in environmental research
- DataONE
- Knowledge Network for Biocomplexity
- eMammal
- Community-based guidelines
- Ask your librarian – don’t be shy!

DataONE Training and Education



Knowledge Network for Biocomplexity

knb ABOUT DATA SHARE TOOLS Jump to: DOI or ID Go SIGN IN

Tools

Over the years, many tools have been developed to facilitate effective data management, archiving content, and retrieving data for synthetic analysis projects.

Morpho

Data management for earth, environmental and ecological scientists.

Morpho allows researchers to create metadata, (i.e. describe their data in a standardized format), and create a catalog of data & metadata upon which to query, edit and view data collections. In addition, it also provides the means to access network servers - like the KNB - in order to query, view and retrieve all relevant, public ecological data! Check the Morpho User Guide for complete details.

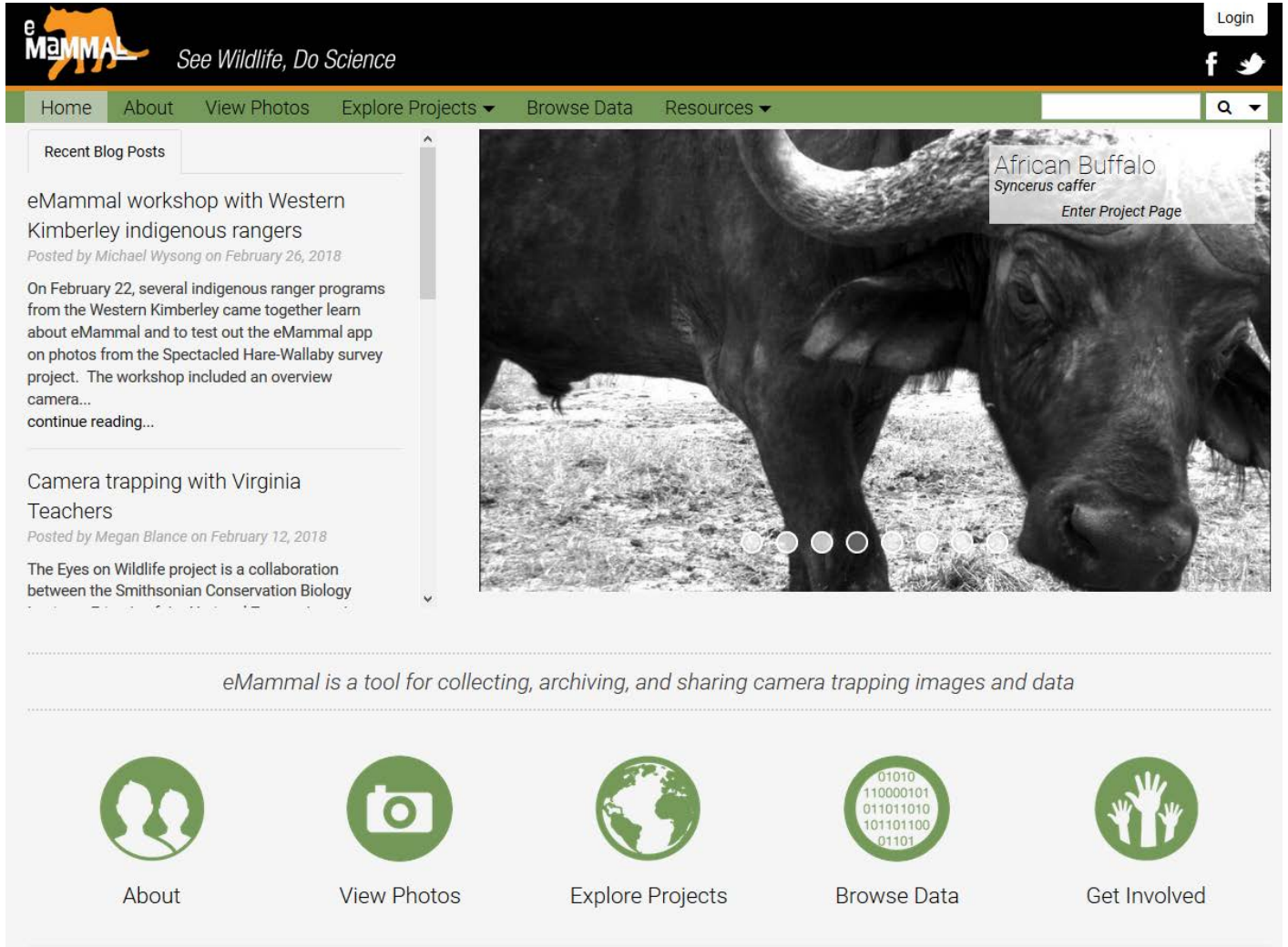
- Create and Edit EML metadata
- Search for existing data packages
- View and download data packages
- Verify and Edit your data
- Specify Access Control rules for your data
- Share and publish your data via the KNB

Download the Morpho data management application.

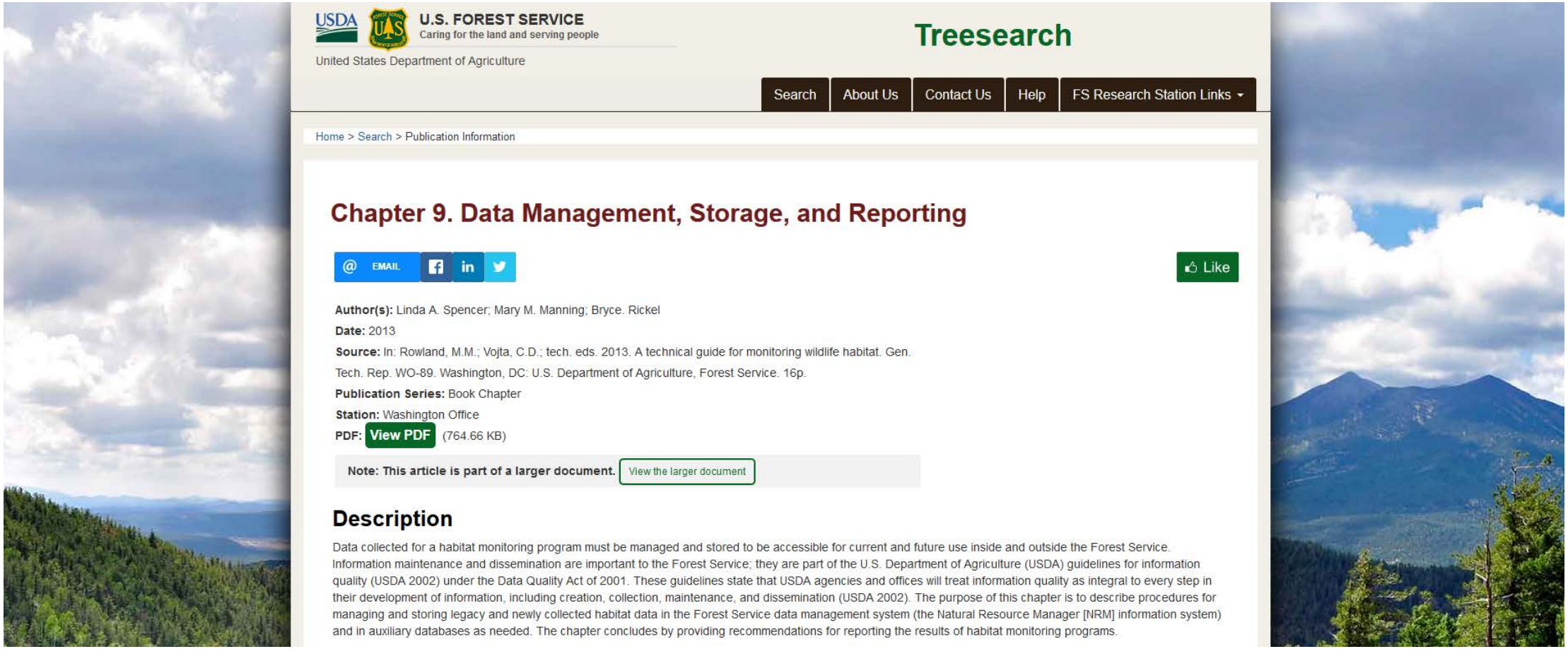
Easy-to-use installers are available for various platforms. For windows and mac versions, users need to double click the installers. For linux version, users need to run "java -jar morpho-version-linux.jar". To run Morpho, you must have Java 1.7 or later installed on your computer.

- Windows
- Linux
- Mac OSX
-
- Older Versions »

eMammal (Smithsonian Institute)



Community best practices in the literature



USDA U.S. FOREST SERVICE
Caring for the land and serving people
United States Department of Agriculture

Treesearch

Search | About Us | Contact Us | Help | FS Research Station Links

Home > Search > Publication Information

Chapter 9. Data Management, Storage, and Reporting

EMAIL | Like

Author(s): Linda A. Spencer; Mary M. Manning; Bryce. Rickel
Date: 2013
Source: In: Rowland, M.M.; Vojta, C.D.; tech. eds. 2013. A technical guide for monitoring wildlife habitat. Gen. Tech. Rep. WO-89. Washington, DC: U.S. Department of Agriculture, Forest Service. 16p.
Publication Series: Book Chapter
Station: Washington Office
PDF: View PDF (764.66 KB)

Note: This article is part of a larger document. View the larger document

Description

Data collected for a habitat monitoring program must be managed and stored to be accessible for current and future use inside and outside the Forest Service. Information maintenance and dissemination are important to the Forest Service; they are part of the U.S. Department of Agriculture (USDA) guidelines for information quality (USDA 2002) under the Data Quality Act of 2001. These guidelines state that USDA agencies and offices will treat information quality as integral to every step in their development of information, including creation, collection, maintenance, and dissemination (USDA 2002). The purpose of this chapter is to describe procedures for managing and storing legacy and newly collected habitat data in the Forest Service data management system (the Natural Resource Manager [NRM] information system) and in auxiliary databases as needed. The chapter concludes by providing recommendations for reporting the results of habitat monitoring programs.

Your turn

- Questions?

Ali Krzton, Research Data Management Librarian
alk0043@auburn.edu
334-844-8268
libguides.auburn.edu/researchdata

References

- Lee, David S. (2012). The Biltmore Forest School: Poking Back into an Extraordinary Time. *The American Biology Teacher 74*:7, 464-469. doi: 10.1525/abt.2012.74.7.7
- Tessarolo, Geziane, Ladle, Richard, Rangel, Thiago, and Hortal, Joaquin. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution 7*:17, 6863-6870. doi: 10.1002/ece3.3259
- Urban, Tim. (2015). The Procrastination Matrix. *Wait But Why*. https://waitbutwhy.com/2015/03/procrastination-matrix.html
- DataONE Education Modules. https://www.dataone.org/education-modules
- Grolemund, Garett, and Wickham, Hadley. (2017). Tidy Data. In: *R for Data Science*. http://r4ds.had.co.nz/tidy-data.html
- MIT Libraries Data Management Workshops. https://libraries.mit.edu/data-management/services/workshops/
- Knowledge Network for Biocomplexity – Tools. https://knb.ecoinformatics.org/#tools
- eMammal. Smithsonian Institute. https://emammal.si.edu
- Spencer, Linda A., Manning, Mary M., and Rickel, Bryce. (2013). Data Management, Storage, and Reporting. In: *A Technical Guide for Monitoring Wildlife Habita*t (tech. eds. Rowland, M.M.; Vojta, C.D.), Gen. Tech. Rep. WO-89. Washington, DC: U.S. Department of Agriculture, Forest Service. 16p. https://www.fs.usda.gov/treesearch/pubs/45226