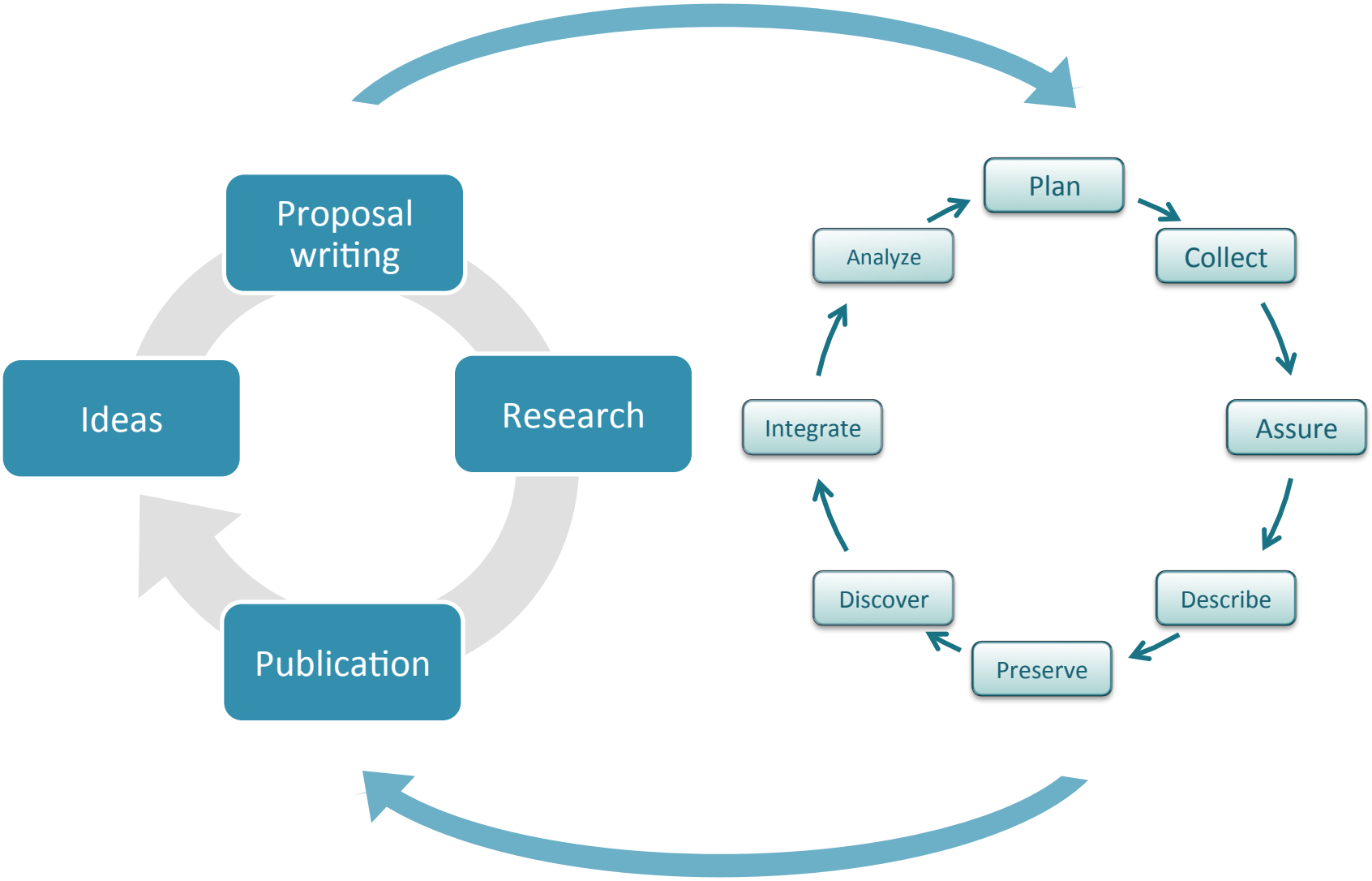
The Data Life Cycle

Bill Michener

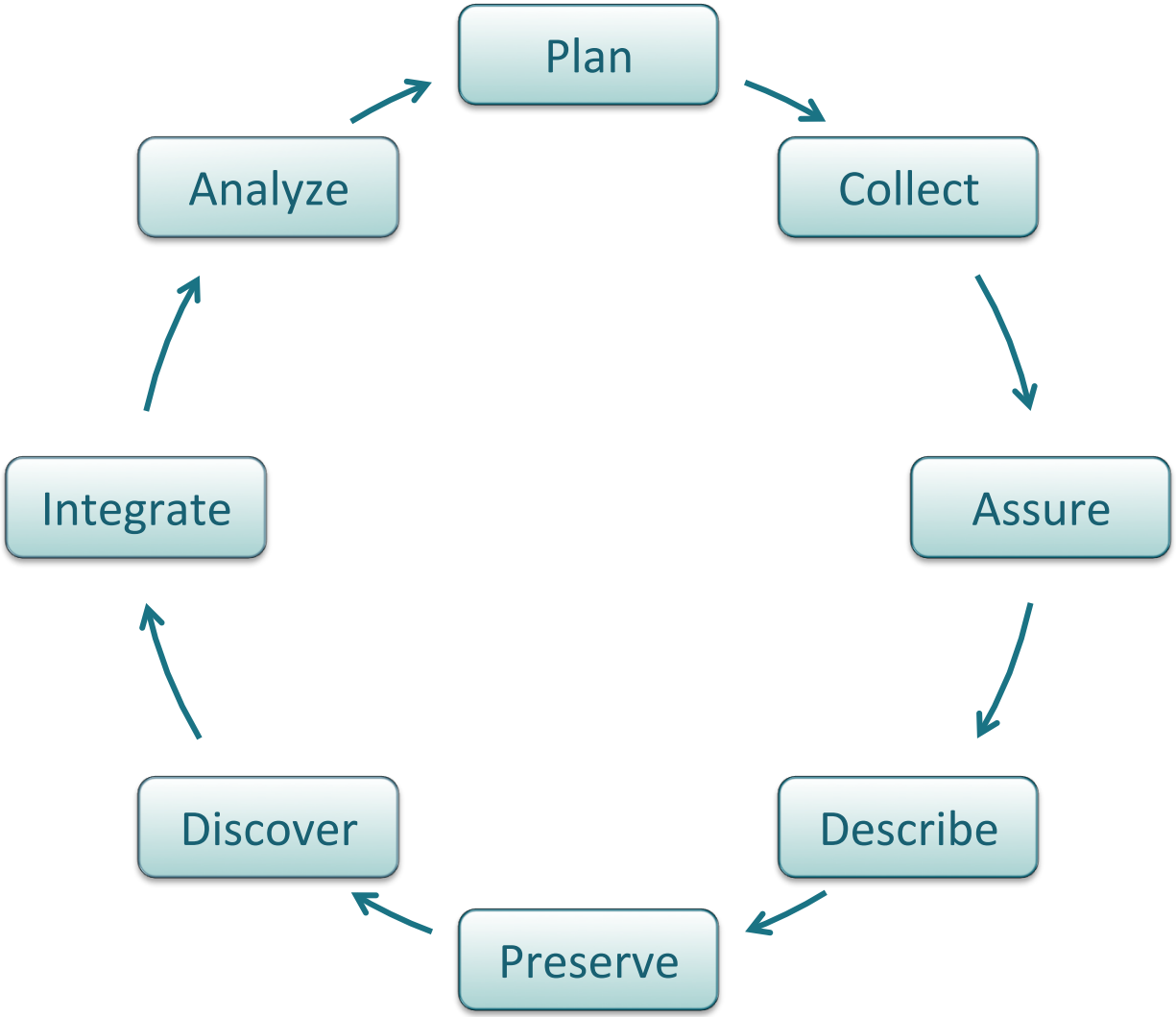
University Libraries, University of New Mexico

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

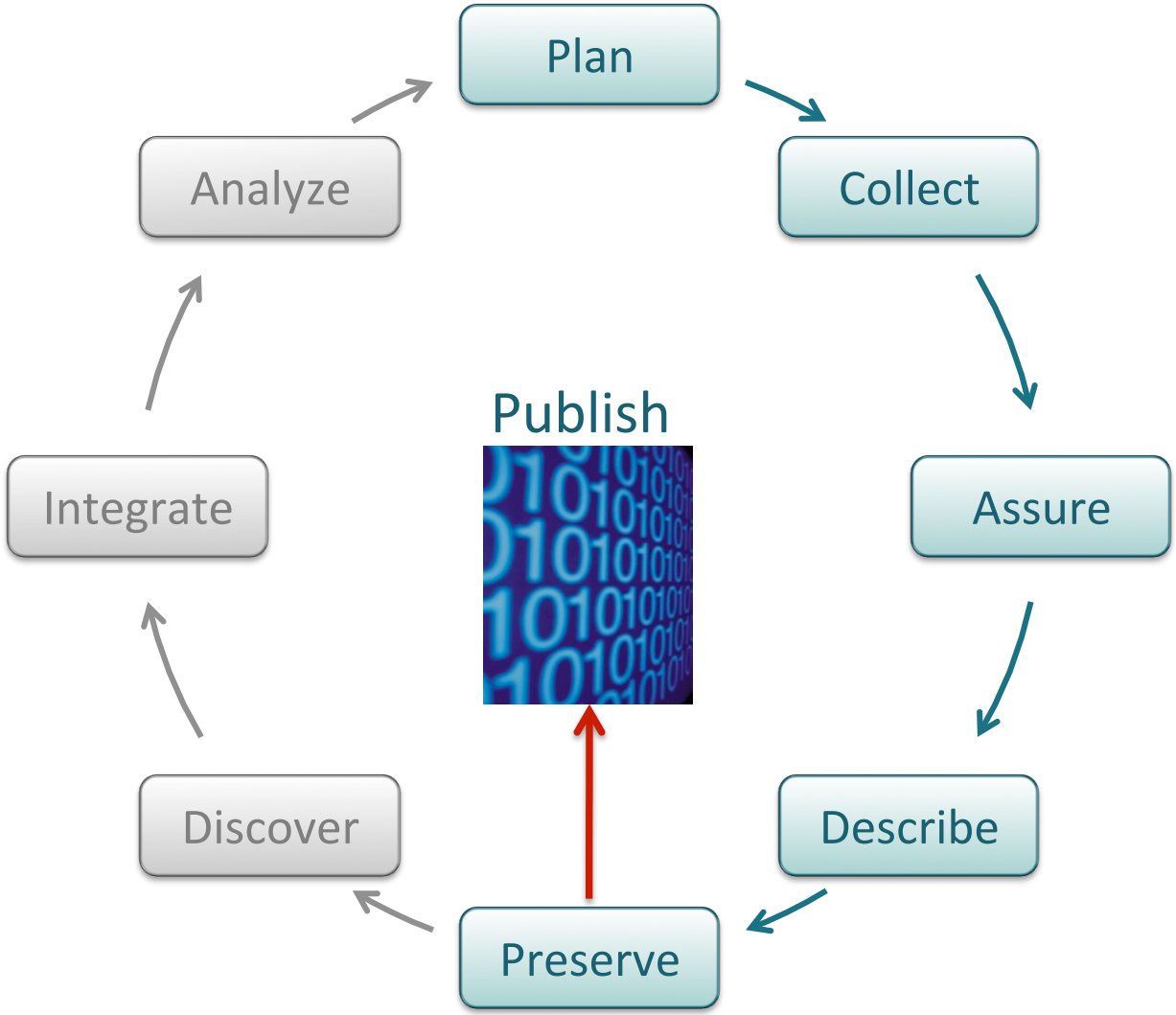
Research and Data Life Cycles



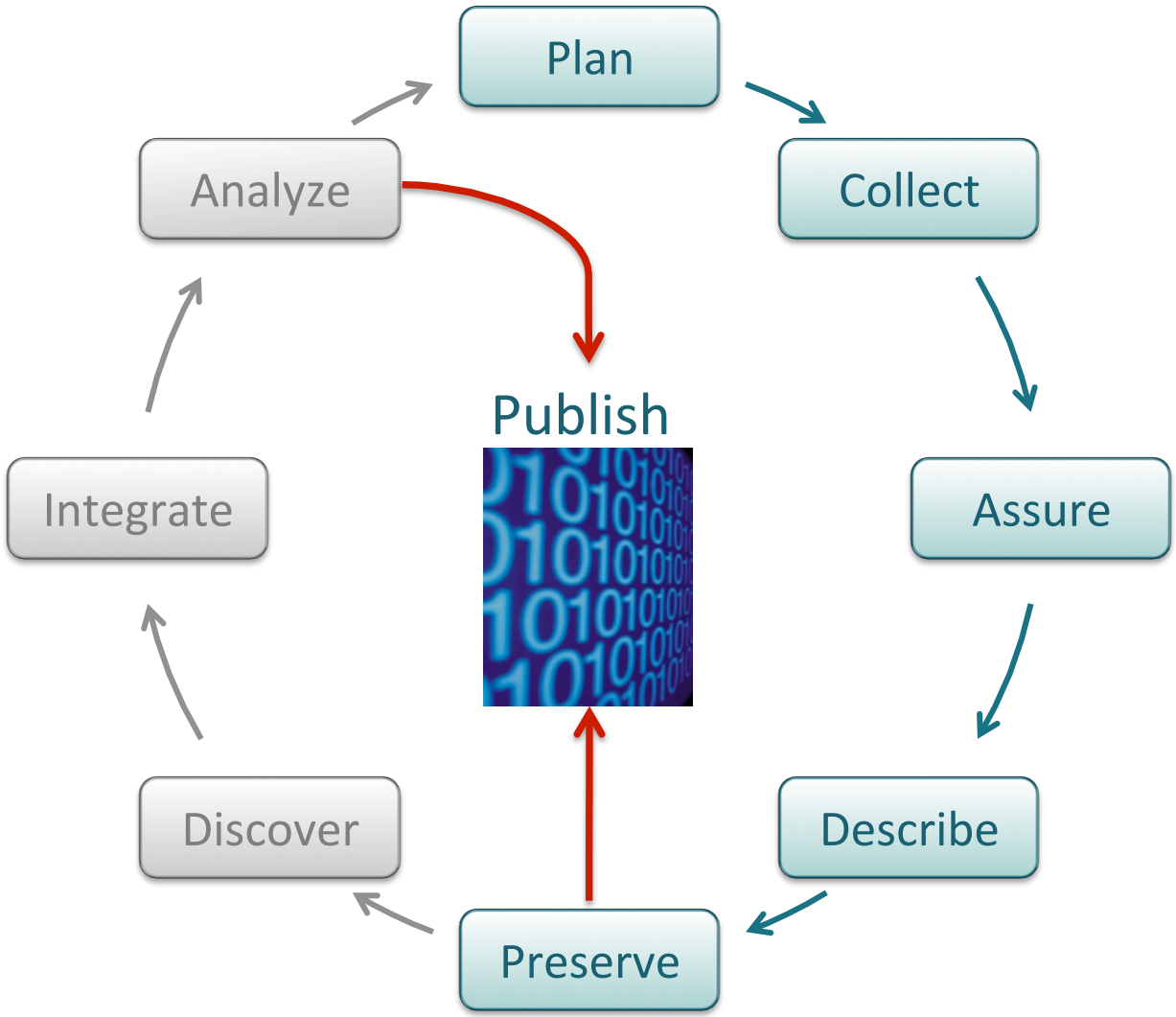
The Data Life Cycle



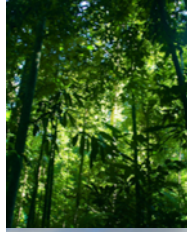
The Data Life Cycle

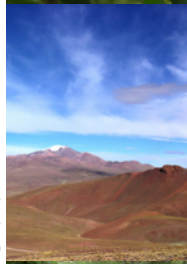
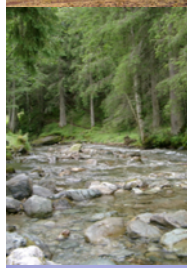


The Data Life Cycle



1. Plan

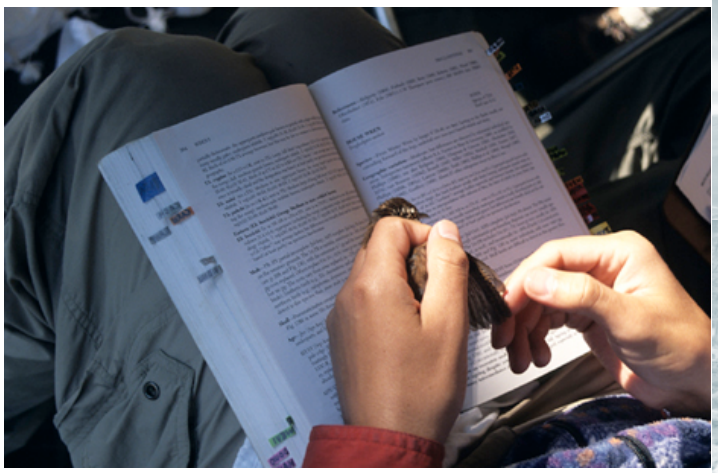


Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

What is a Data Management Plan?

- A document that describes what you will do with your data **during** and **after** you complete your research



2. Collect (e.g., organize data)

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

Organizing data: Spreadsheets:
primordial data entry tool of the digital age

	A	B	C	D	
1			Trapping Intervals		
2		Trapping dates and times			
3		Down		Up	
4	Date	Time	Up	Time	Tim
5	2/8/10	16:30	2/9/10	8:30	
6	2/9/10	8:30	2/9/10	17:45	
8	2/10/10	7:45	2/10/10	17:05	
9	2/10/10	17:05	2/11/10	7:25	
10	2/11/10	7:30	2/11/10	17:35	
12	2/12/10	7:20	2/12/10	17:45	
13	2/12/10	17:45	2/13/10	7:45	
14	2/13/10	7:45	2/13/10	17:50	
15	2/13/10	17:50	2/14/10	7:30	
16	2/14/10	7:30	2/14/10	17:50	
17	2/14/10	17:50	2/15/10	7:20	
18	2/15/10	7:20	2/15/10	18:00	
19	2/15/10	18:00	2/16/10	8:20	
20	2/16/10	8:25	2/16/10	17:10	
21	2/16/10	17:10	2/17/10	8:05	
22	2/17/10	8:10	2/17/10	16:55	
24	2/18/10	7:25	2/18/10	17:45	

Best Practices

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	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

SEY_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

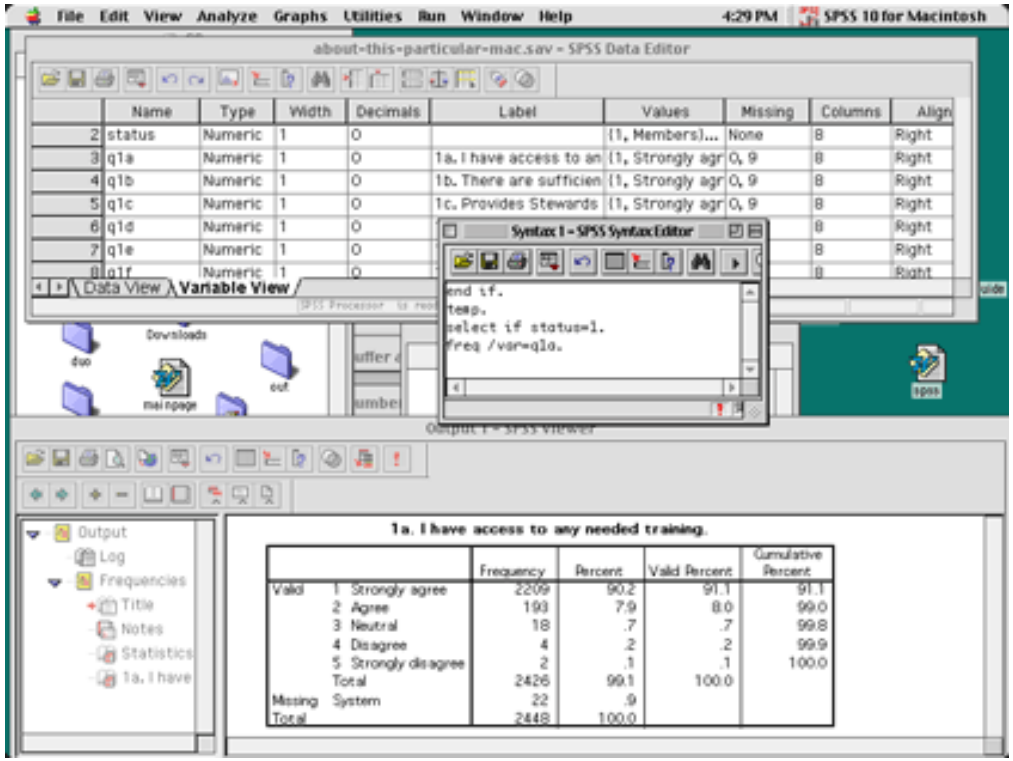
- Columns of data are consistent:
  - only numbers, dates, or text
- Consistent Names, Codes, Formats (date) used in each column
- Data are all in one table – computer works better on single table than multiple small tables that require a lot of human intervention
- Descriptive File Name

Statistical Software

R, SAS, MATLAB, SPSS,

You can implement good data practices in each of these…

Of course, these are also excellent for calculations, data analysis, quality assurance, sub-setting data



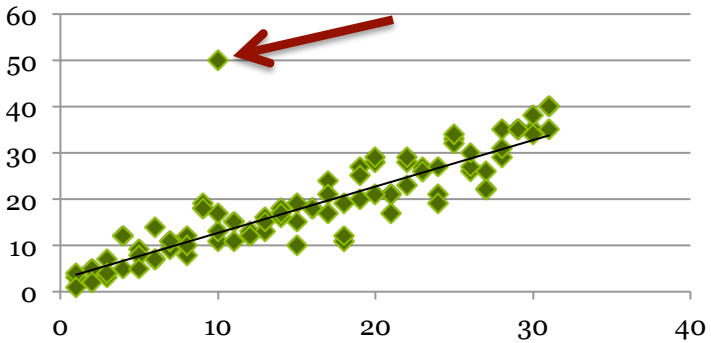
3. Assure

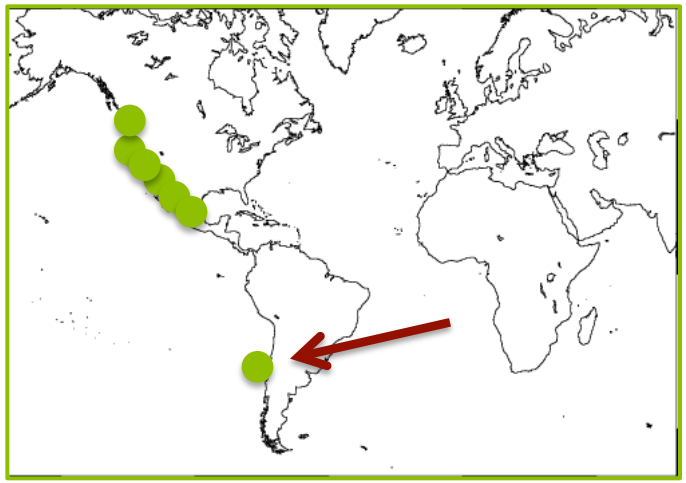
Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

QA/QC Approaches

- Ensure data line up in columns
- No missing, impossible, or anomalous values

  sort by fields to highlight discrepancies
- Perform statistical and graphical summaries





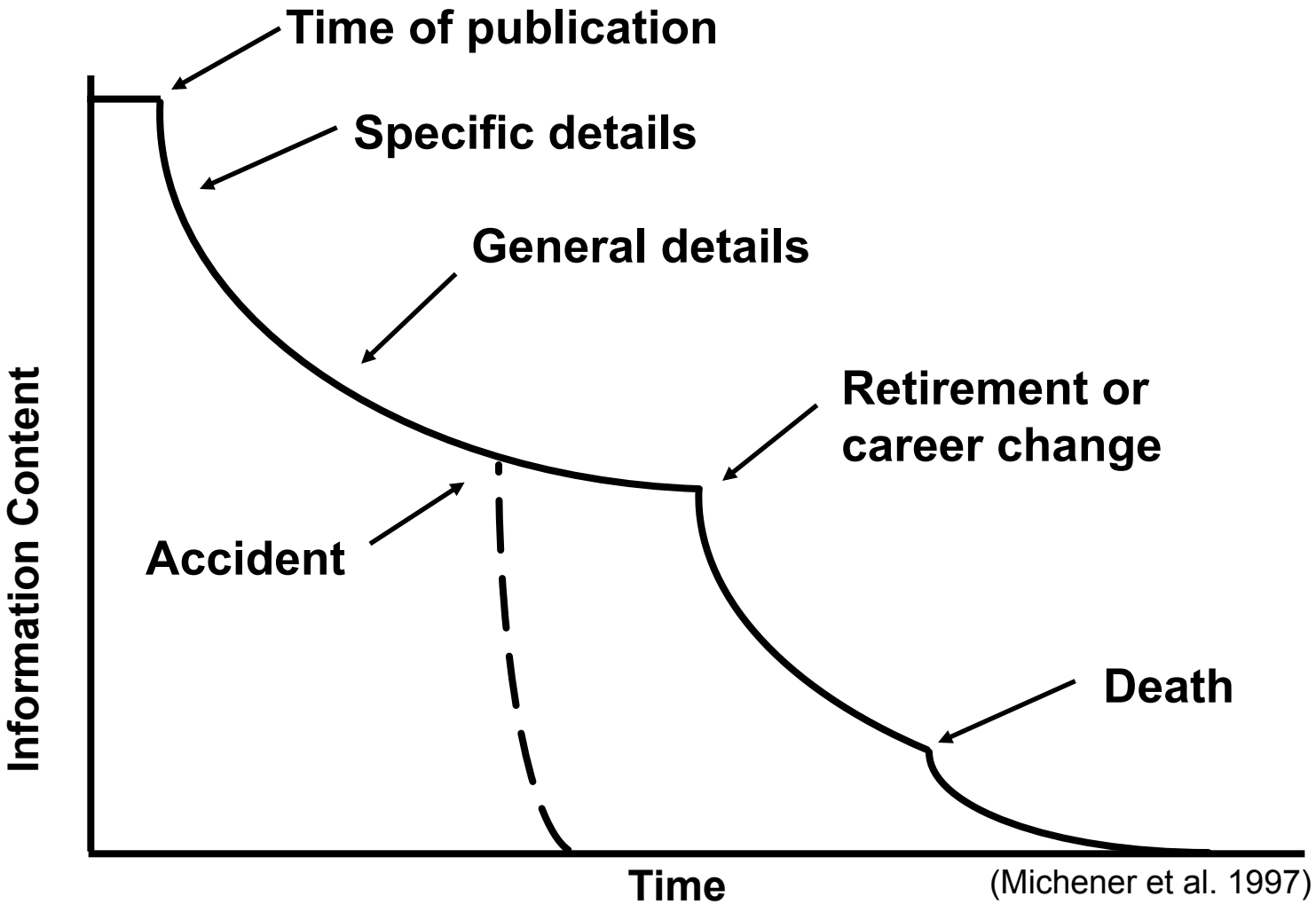
4. Describe

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

Metadata = description of data

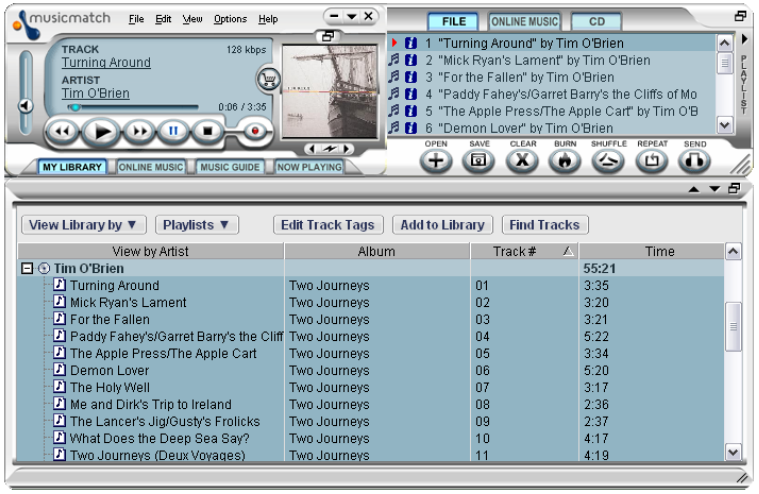
- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** was it created?
- **WHERE** is it geographically?
- **HOW** was the data processed?
- **WHY** was the data set developed?

Why metadata? Data entropy

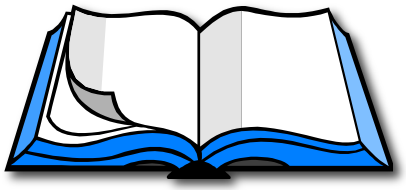


(Michener et al. 1997)

Metadata in the Real World



Metadata is all around…



Author(s)	Boullosa, Carmen.
Title(s)	They're cows, we're pigs / by Carmen Boullosa
Place	New York : Grove Press, 1997.
Physical Descr	viii, 180 p ; 22 cm.
Subject(s)	Pirates Caribbean Area Fiction.
Format	Fiction

Multiple metadata standards exist

Darwin Core

- Emphasis on museum specimens

Dublin Core Element Set

- Emphasis on web resources, publications

Ecological Metadata Language (EML)

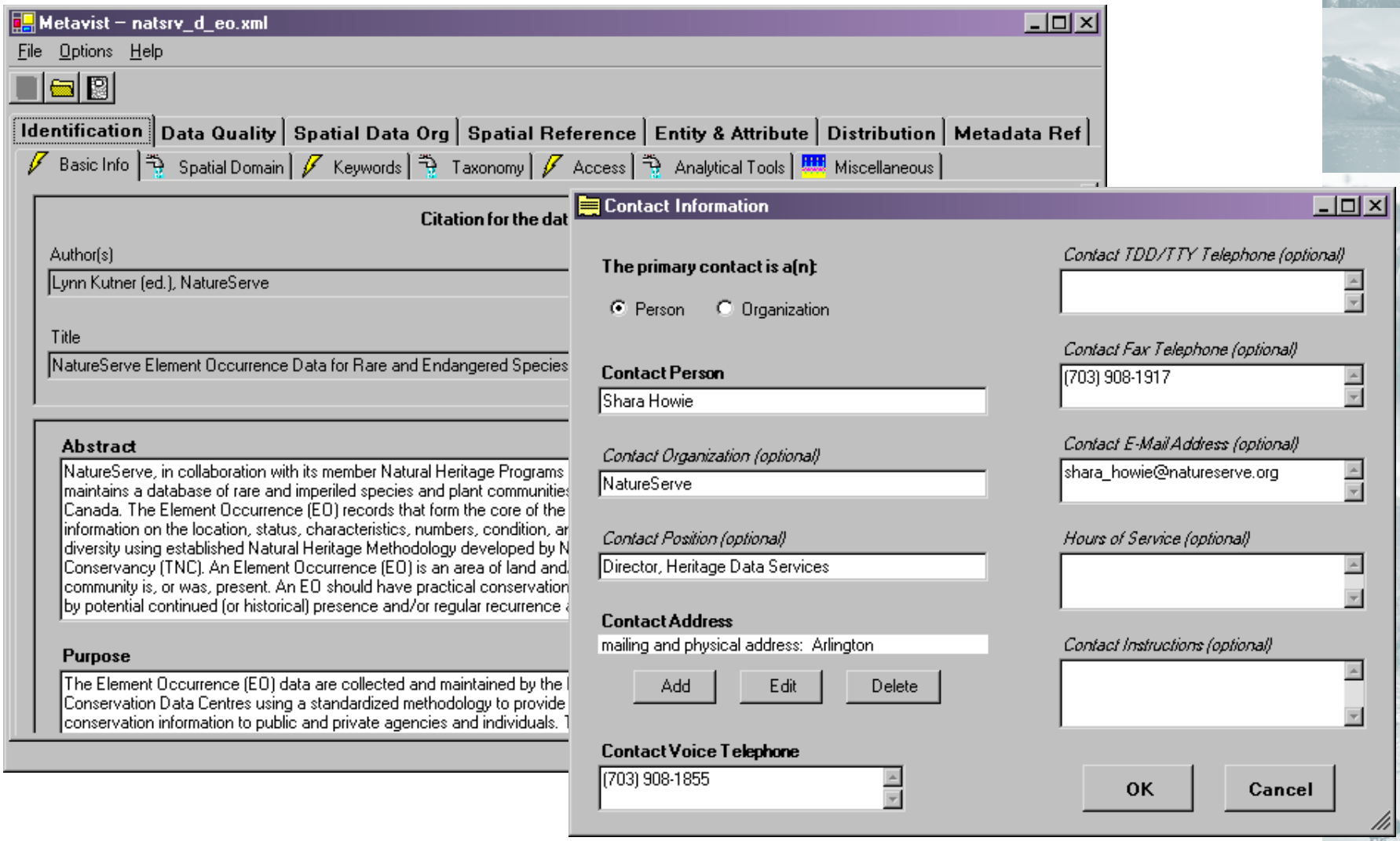
- Emphasis on ecology and environmental sciences

ISO 19115 Geographic information: Metadata

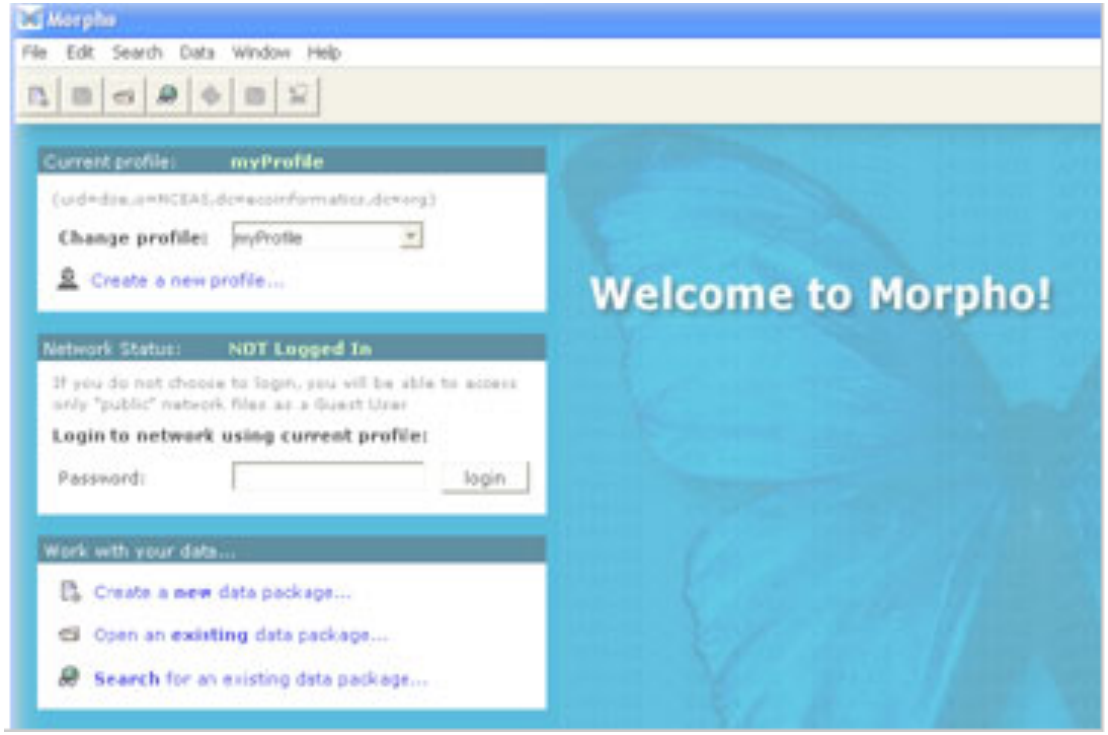
- Emphasis on geospatial data and services

+ dozens of others…

Tools: Metavist (geospatial data)



Morpho: An Ecological Metadata Language editor



5. Preserve

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

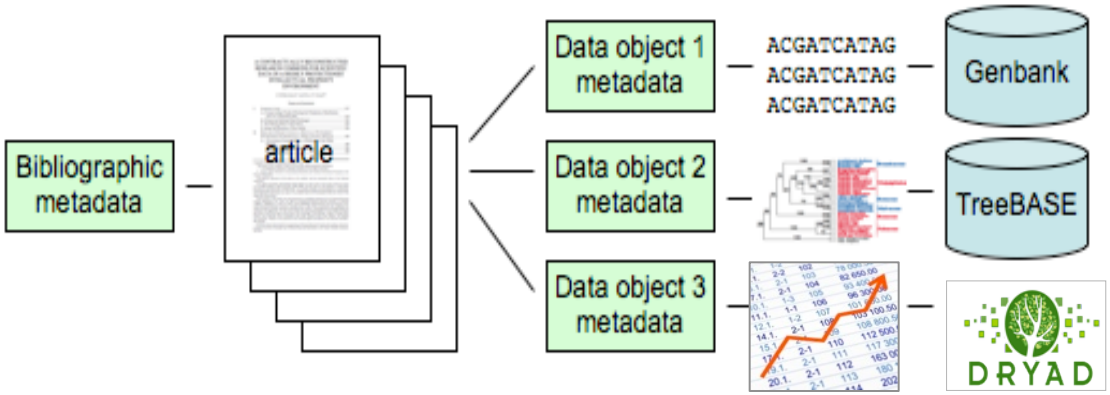
Data Center: Stewardship and Archive Functions

- Acquisition
  - identify how best to serve the scientific community
  - establish how and when to receive data
- Ingest
  - perform QA checks
  - compile project-provided metadata
  - convert to archivable file formats
- Enhance (as requested)
  - convert to standard formats & units
  - aggregate files
- Metadata / Documentation
  - Prepare final metadata record and documentation
- Archive / Publish
  - generate citation
- Exploration and Distribution
  - provide tools to explore, access, and extract data for users worldwide
- Post-Project Data Support
  - serve as a buffer between end users and PIs
  - provide usage statistics
- Stewardship
  - provide long-term secure archiving of the data
  - security, disaster recovery
  - migration to new computer systems

Dryad (~3,000 data products)

Coordinated submission of articles and underlying data

Handshaking with specialized repositories

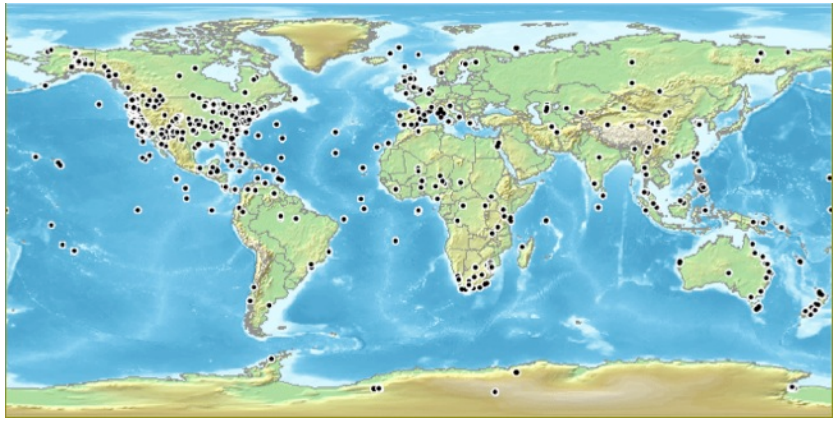


Promotion of reuse and incentives for deposit

CC 0 ZERO

DataCite
International Data Citation

Knowledge Network for Biocomplexity (> 50,000 data packages)



Data Types

- Ecological
- Environmental
- Demographic
- Social/Legal/Economic

Contributors

- Individual investigators
- Field stations and networks
- Government agencies
- Non-profit partnerships
- Synthesis centers

Additional Data Centers

AmeriFlux

ORNL DAAC Distributed Active Archive Center for Biogeochemical Dynamics

CUAHSI HIS Sharing hydrologic data

BCO-DMO Biological and Chemical Oceanography Data Management Office

GLOBAL BIODIVERSITY INFORMATION FACILITY

NEODC

TreeBASE

Genbank

CDIAC

USGS Earth Science Data Clearinghouse Formerly nbii

6. Discover

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

Search does not equal discovery

Google

soil organic carbon

Search

About 1,700,000 results (0.29 seconds)

Web
Images
Maps
Videos
News
Shopping
More

Santa Fe, NM
Change location

Show search tools

Ad related to **soil organic carbon** Why this ad?

Organic Soil | Lowes.com
www.lowes.com/ - ★★★★☆ 168 seller reviews
Find Gardening Supplies At Lowe's® Official Site Today. Shop Now!
1,471 people +1'd Lowe's Home Improvement
+ Show map of 3458 Zafarano Dr, Santa Fe, NM

Scholarly articles for soil organic carbon
Total **carbon, organic carbon,** and **organic** matter. - Nelson - Cited by 6671
Soil carbon pools and world life zones - Post - Cited by 1428
The vertical distribution of **soil organic carbon** and its ... - Jobbágy - Cited by 935

Soil carbon - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Soil_carbon**
Soil organic matter, of which carbon is a major part, holds a great proportion of ...
Tillage and drainage both expose **soil organic matter** to oxygen and oxidation.
↳ Overview - Soil carbon and soil health - Losses of soil carbon

Soil organic carbon
www.eoearth.org/article/**Soil_organic_carbon**
by E Milne - Related articles
Dec 21, 2009 – **Soils** contain **carbon** (C) in both **organic** and inorganic forms. In most **soils** (with the exception of calcareous **soils**) the majority of C is held as ...

[PDF] **The importance of soil organic matter**
ftp://ftp.fao.org/agl/agll/docs/sb80e.pdf
File Format: PDF/Adobe Acrobat - Quick View
by VT di Caracalla - Related articles
Human interventions that influence **soil organic matter**. 15. Practices that ... Evaluation of the organic matter content of a soil in Paraná. 20. 9. Reduction of dry ...

DataONE

About | Participate | Resources | Education | Data

ONEMercury *A DataONE Search Tool for Scientific Data*

Search For:

Hint: boolean operators and phrases are allowed. ex: precipitation or (rain and "moisture content")

Results/Page 10 SEARCH

Show/Hide Advanced Options | Help

Fielded Search

FullText | OR
FullText | OR
FullText

Help | clear

Date Search

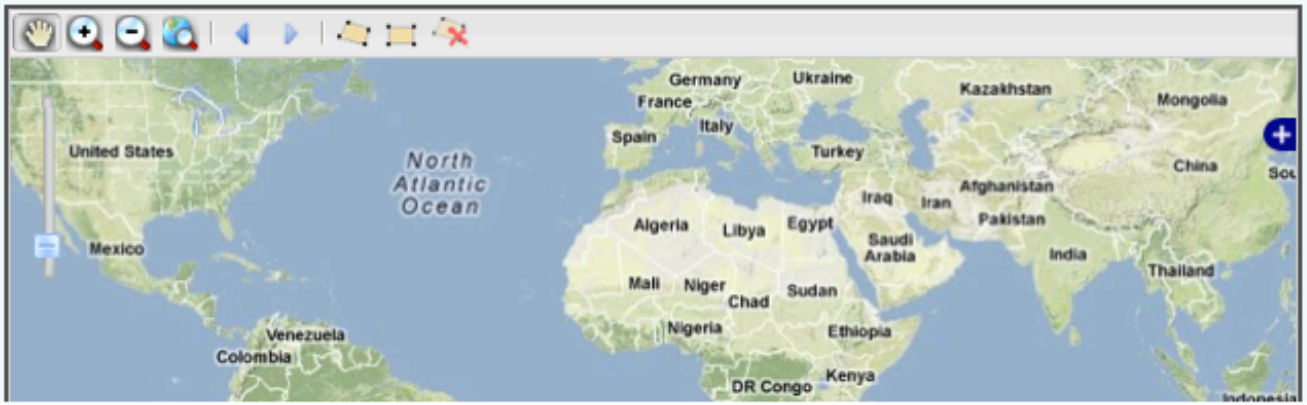
Collection Date / Publication Date / Either

during thru

mm/dd/yyyy *mm/dd/yyyy*

Help | clear

Geographic Search



List Areas in:

USA / WORLD

Select from list

Search Area:

overlaps / encloses

North

ONEMercury *A DataONE Search Tool for Scientific Data*

Return to Search | Email Query | Bookmark Query | RSS Feed for Query | Help

Query: text : water

Hide Filters

Filter by author	Filter by project	Filter by keywords	Filter by Originator
Bruce Menge (6459)	Partnership for Interdisciplinary (14809)	temperature (15631)	Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777)
Margaret McManus (3465)	Florida Coastal Everglades (512)	Oceanographic Sensor Data (15359)	PISCO (1309)
Libe Washburn (3112)	Olympic Coast National (510)	continental shelf (15359)	Monterey Bay National Marine Sanctuary (952)
Jack Barth (1612)	Arctic Long-Term Ecological (339)	Temperature (15275)	Olympic Coast National Marine
Mary Sue Brancato (510)		United States of America (15146)	
Anne Giblin (298)			
Pete Raimondi (237)			

Sort By: Relevance | Date | Member Node

Viewing Documents 1 - 10 out of 22815
Prev 1 2 3 4 5 6 7 8 9 10 Next

Filter by Member Node

- PISCO MN (15460)
- LTER Network Member Node (4289)
- Merritt D1 Member Node (1845)
- Knowledge Network for Biocomplexity (764)
- ORNL DAAC (378)
- SANParks Data Repository (42)
- ESA Data Registry (21)
- USGS Core Sciences Clearinghouse (16)

THE EFFECT OF OVERLAPPING PIOSPHERES ON LANDSCAPE HETEROGENEITY 02/09/2004 - 05/09/2004

Datasource: SANPARKS DATA REPOSITORY

In 1933, Kruger National Park implemented artificial sources of surface water. Many studies have been conducted on the effect these waterholes have on herbivore distribution and the related impacts. One such finding is that piospheres, patches created by herbivores through their grazing, browsing and trampling activities focusing around a water source (Owen-Smith 1996 cited in Gaylard et al 2002) occur around waterholes and contribute to the patchiness in the landscape. The aim of this study was to determine if the proportion of increaser II grass species would drop below 50% and be replaced b...

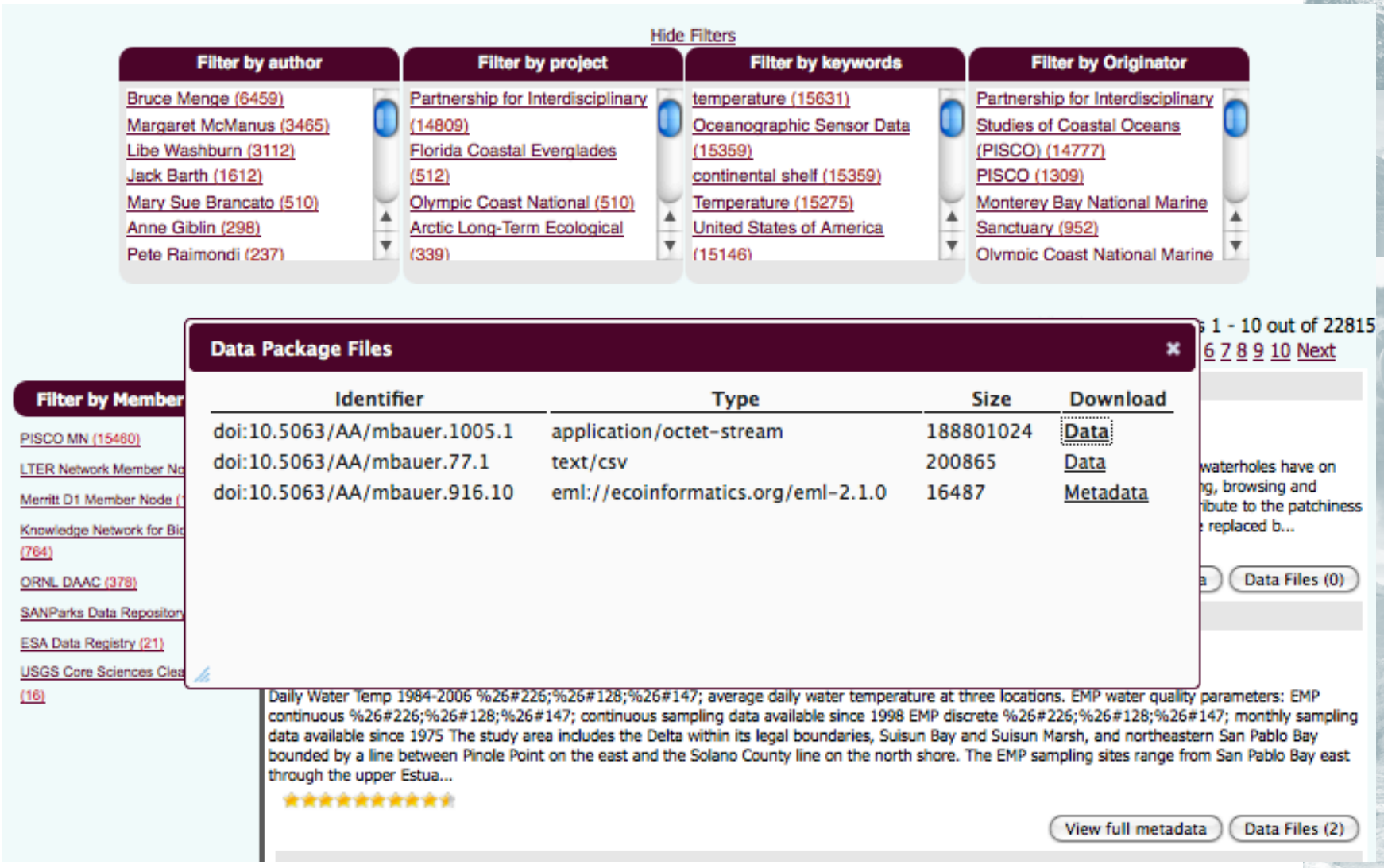
View full metadata | Data Files (0)

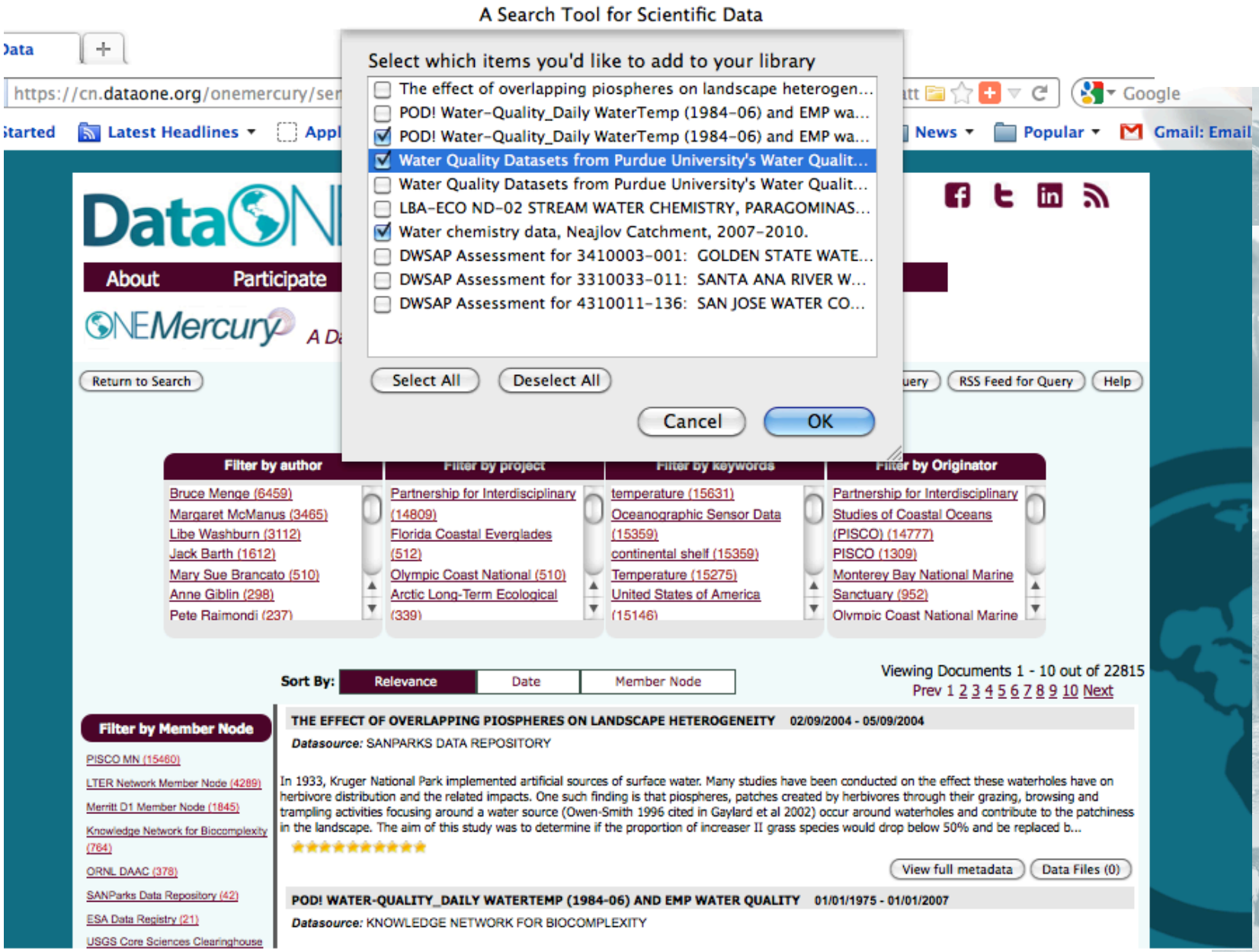
POD! WATER-QUALITY_DAILY WATERTEMP (1984-06) AND EMP WATER QUALITY 01/01/1975 - 01/01/2007

Datasource: KNOWLEDGE NETWORK FOR BIOCOMPLEXITY

Daily Water Temp 1984-2006 %26#226;%26#128;%26#147; average daily water temperature at three locations. EMP water quality parameters: EMP continuous %26#226;%26#128;%26#147; continuous sampling data available since 1998 EMP discrete %26#226;%26#128;%26#147; monthly sampling data available since 1975 The study area includes the Delta within its legal boundaries, Suisun Bay and Suisun Marsh, and northeastern San Pablo Bay bounded by a line between Pinole Point on the east and the Solano County line on the north shore. The EMP sampling sites range from San Pablo Bay east through the upper Estua...

View full metadata | Data Files (2)


Hide Filters
Filter by author
Bruce Menge (6459)
Margaret McManus (3465)
Libe Washburn (3112)
Jack Barth (1612)
Mary Sue Brancato (510)
Anne Giblin (298)
Pete Raimondi (237)
Filter by project
Partnership for Interdisciplinary (14809)
Florida Coastal Everglades (512)
Olympic Coast National (510)
Arctic Long-Term Ecological (339)
Filter by keywords
temperature (15631)
Oceanographic Sensor Data (15359)
continental shelf (15359)
Temperature (15275)
United States of America (15146)
Filter by Originator
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777)
PISCO (1309)
Monterey Bay National Marine Sanctuary (952)
Olympic Coast National Marine
Data Package Files
Identifier
Type
Size
Download
doi:10.5063/AA/mbauer.1005.1
application/octet-stream
188801024
Data
doi:10.5063/AA/mbauer.77.1
text/csv
200865
Data
doi:10.5063/AA/mbauer.916.10
eml://ecoinformatics.org/eml-2.1.0
16487
Metadata
1 - 10 out of 22815
6 7 8 9 10 Next
Filter by Member
PISCO MN (15460)
LTER Network Member No
Merritt D1 Member Node (
Knowledge Network for Bio (764)
ORNL DAAC (378)
SANParks Data Repository
ESA Data Registry (21)
USGS Core Sciences Clea (16)
waterholes have on
ng, browsing and
ibute to the patchiness
replaced b...
Data Files (0)
Daily Water Temp 1984-2006 %26#226;%26#128;%26#147; average daily water temperature at three locations. EMP water quality parameters: EMP continuous %26#226;%26#128;%26#147; continuous sampling data available since 1998 EMP discrete %26#226;%26#128;%26#147; monthly sampling data available since 1975 The study area includes the Delta within its legal boundaries, Suisun Bay and Suisun Marsh, and northeastern San Pablo Bay bounded by a line between Pinole Point on the east and the Solano County line on the north shore. The EMP sampling sites range from San Pablo Bay east through the upper Estua...
View full metadata
Data Files (2)


A Search Tool for Scientific Data
Select which items you'd like to add to your library
The effect of overlapping piospheres on landscape heterogen...
POD! Water-Quality_Daily WaterTemp (1984-06) and EMP wa...
POD! Water-Quality_Daily WaterTemp (1984-06) and EMP wa...
Water Quality Datasets from Purdue University's Water Qualit...
Water Quality Datasets from Purdue University's Water Qualit...
LBA-ECO ND-02 STREAM WATER CHEMISTRY, PARAGOMINAS...
Water chemistry data, Neajlov Catchment, 2007-2010.
DWSAP Assessment for 3410003-001: GOLDEN STATE WATE...
DWSAP Assessment for 3310033-011: SANTA ANA RIVER W...
DWSAP Assessment for 4310011-136: SAN JOSE WATER CO...
Select All
Deselect All
Cancel
OK
About
Participate
Return to Search
RSS Feed for Query
Help
Filter by author
Bruce Menge (6459)
Margaret McManus (3465)
Libe Washburn (3112)
Jack Barth (1612)
Mary Sue Brancato (510)
Anne Giblin (298)
Pete Raimondi (237)
Partnership for Interdisciplinary (14809)
Florida Coastal Everglades (512)
Olympic Coast National (510)
Arctic Long-Term Ecological (339)
temperature (15631)
Oceanographic Sensor Data (15359)
continental shelf (15359)
Temperature (15275)
United States of America (15146)
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777)
PISCO (1309)
Monterey Bay National Marine Sanctuary (952)
Olympic Coast National Marine
Sort By:
Relevance
Date
Member Node
Viewing Documents 1 - 10 out of 22815
Prev 1 2 3 4 5 6 7 8 9 10 Next
Filter by Member Node
PISCO MN (15460)
LTER Network Member Node (4289)
Merritt D1 Member Node (1845)
Knowledge Network for Biocomplexity (764)
ORNL DAAC (378)
SANParks Data Repository (42)
ESA Data Registry (21)
USGS Core Sciences Clearinghouse
THE EFFECT OF OVERLAPPING PIOSPHERES ON LANDSCAPE HETEROGENEITY 02/09/2004 - 05/09/2004
Datasource: SANPARKS DATA REPOSITORY
In 1933, Kruger National Park implemented artificial sources of surface water. Many studies have been conducted on the effect these waterholes have on herbivore distribution and the related impacts. One such finding is that piospheres, patches created by herbivores through their grazing, browsing and trampling activities focusing around a water source (Owen-Smith 1996 cited in Gaylard et al 2002) occur around waterholes and contribute to the patchiness in the landscape. The aim of this study was to determine if the proportion of increaser II grass species would drop below 50% and be replaced b...
View full metadata
Data Files (0)
POD! WATER-QUALITY_DAILY WATERTEMP (1984-06) AND EMP WATER QUALITY 01/01/1975 - 01/01/2007
Datasource: KNOWLEDGE NETWORK FOR BIOCOMPLEXITY

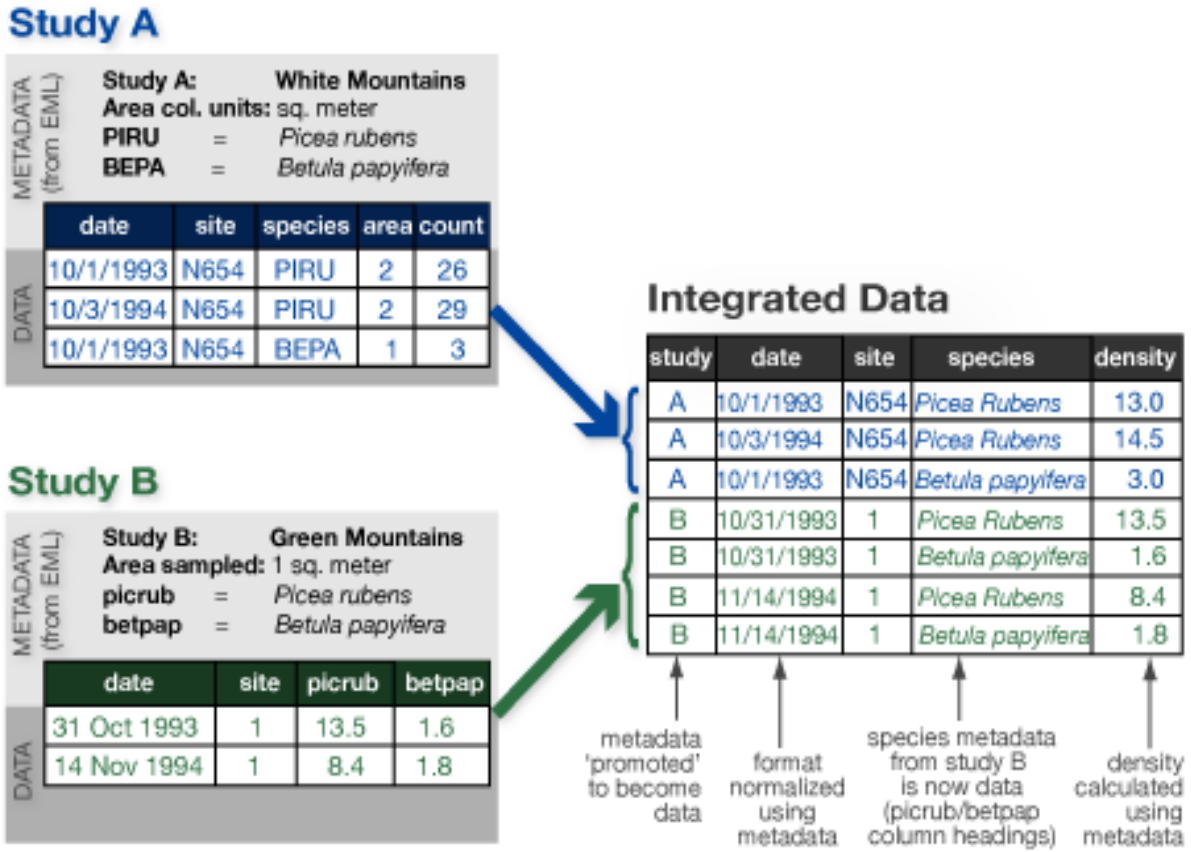
7. Integrate

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

Data integration challenges

Heterogeneity:

Syntax
(format)
Schema
(model)
Semantics
(meaning)

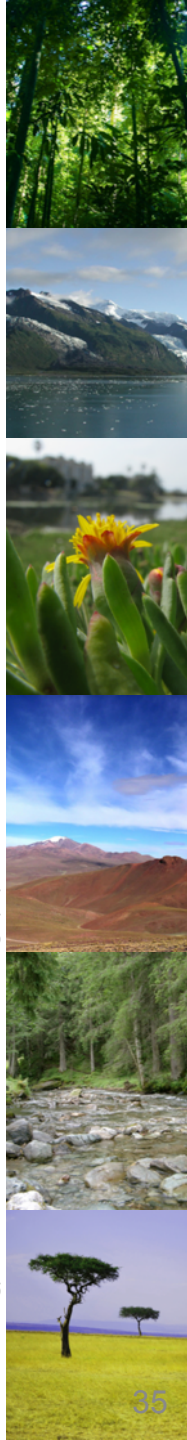


Jones et al. 2007

Integration best practices

- Review metadata for project context, sampling methods (time and space), QA/QC, etc. (i.e., adequacy and relevance—can the data be used?)
- Examine data format and align parameters (if possible?)
  - Date/time, geospatial formats
- Merge using appropriate software (e.g., R, SAS, etc.)
- Verify (use QA/QC approaches and visualization techniques)
  - Range checks, illegal value filters, etc.
  - Exploratory stats (stem-leaf plots, X-Y plots, etc.)
  - Check data summaries

8. Analyze

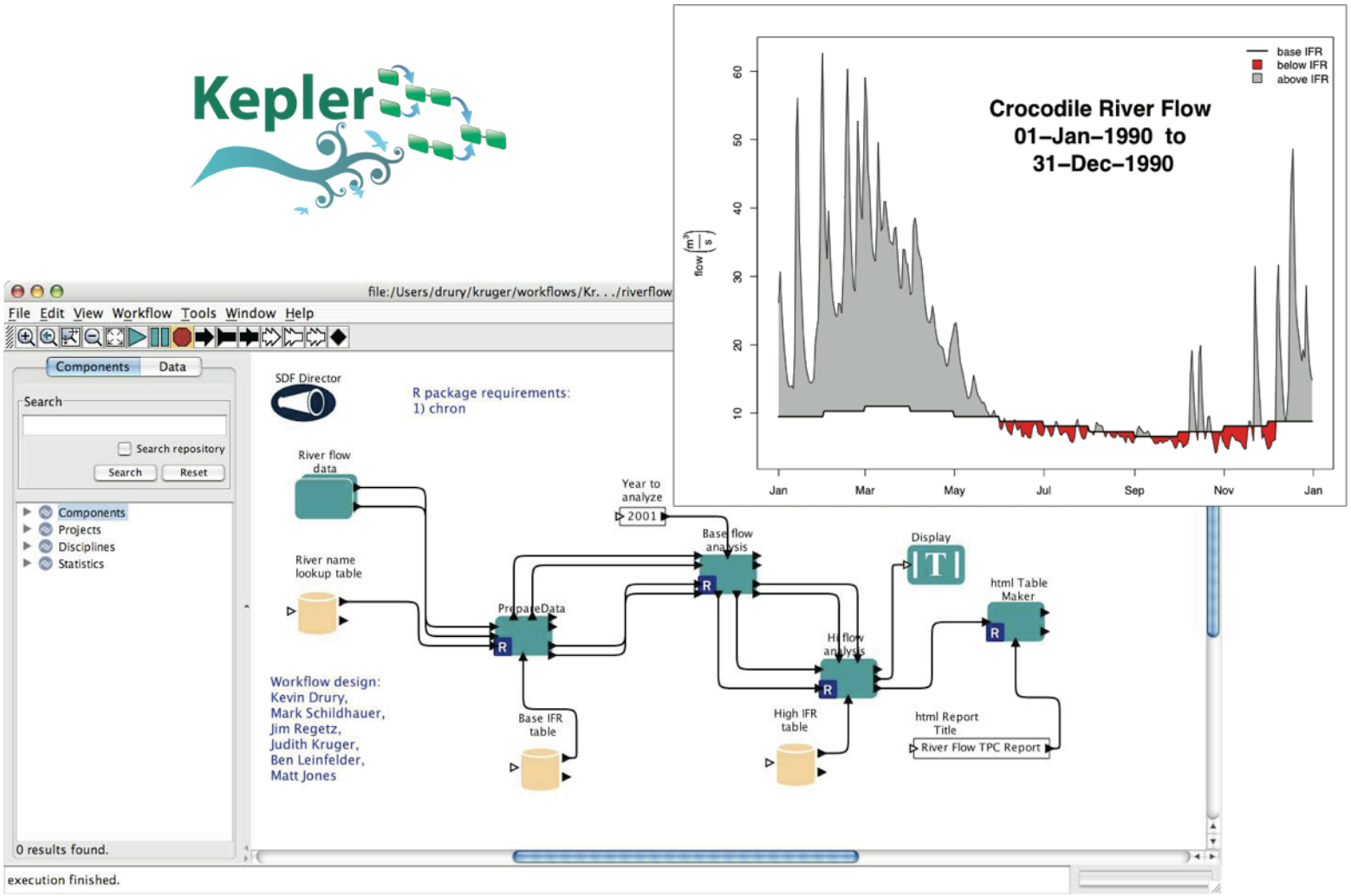


Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

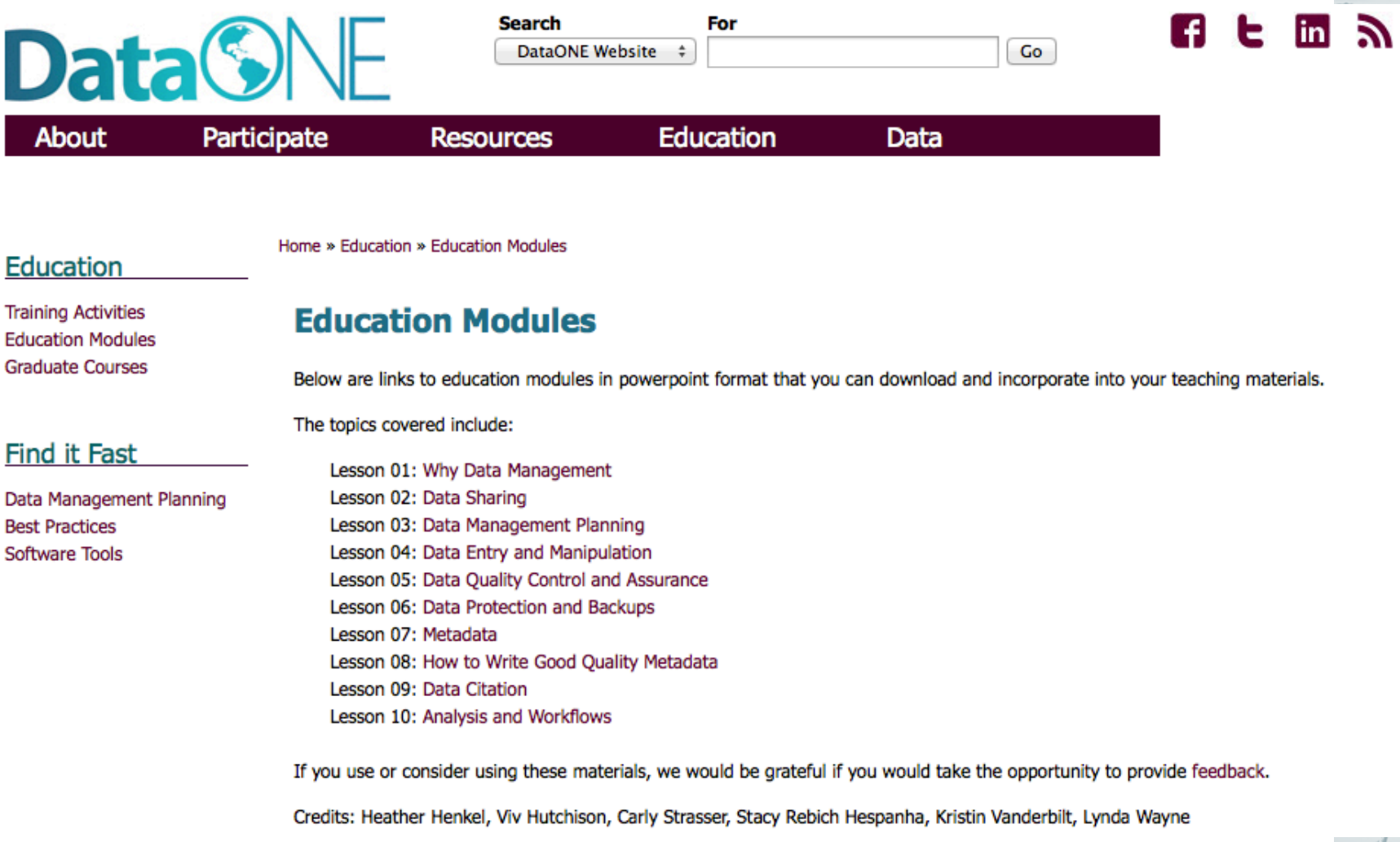
Analysis tools

1. Programming languages
   a) C, C++
   b) Fortran
   c) Perl
   d) Python
2. Statistics and Analysis
   a) Excel
   b) R
   c) MATLAB
   d) SAS
   e) SPSS
3. Scientific workflows
   a) Kepler
   b) myExperiment
   c) Pegasus
   d) Taverna
   e) VisTrails

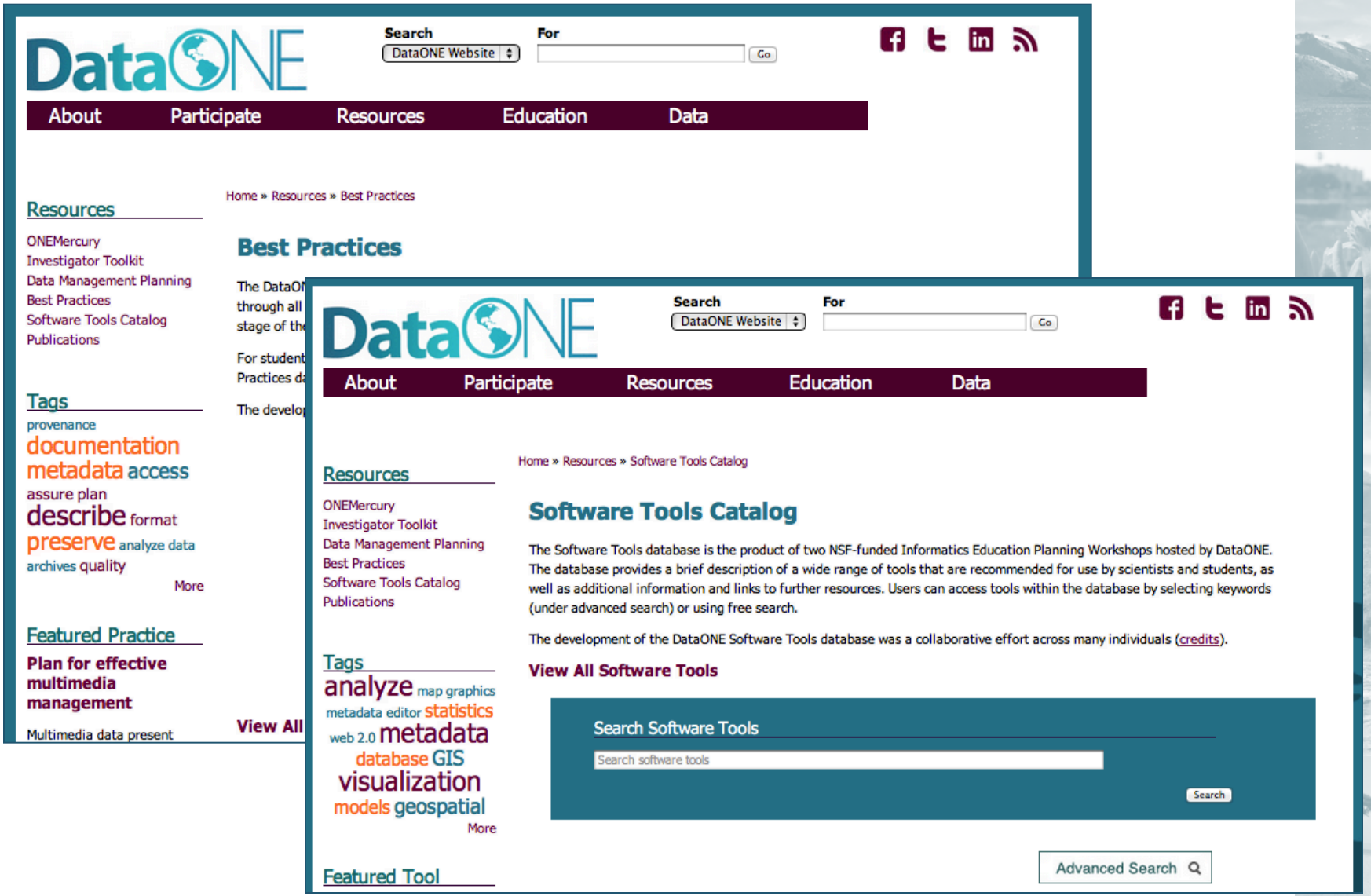
Scientific workflows



Find out more: DataONE.org



Best Practices and Software Tools



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

Thanks!

Bob Cook, Amber Budden, Patricia Cruse, Don Edwards, Viv Hutchison, Matt Jones, Steve Kelling, Andrew Sallans, Mark Schildhauer, Carly Strasser, Kristin Vanderbilt, and Todd Vision