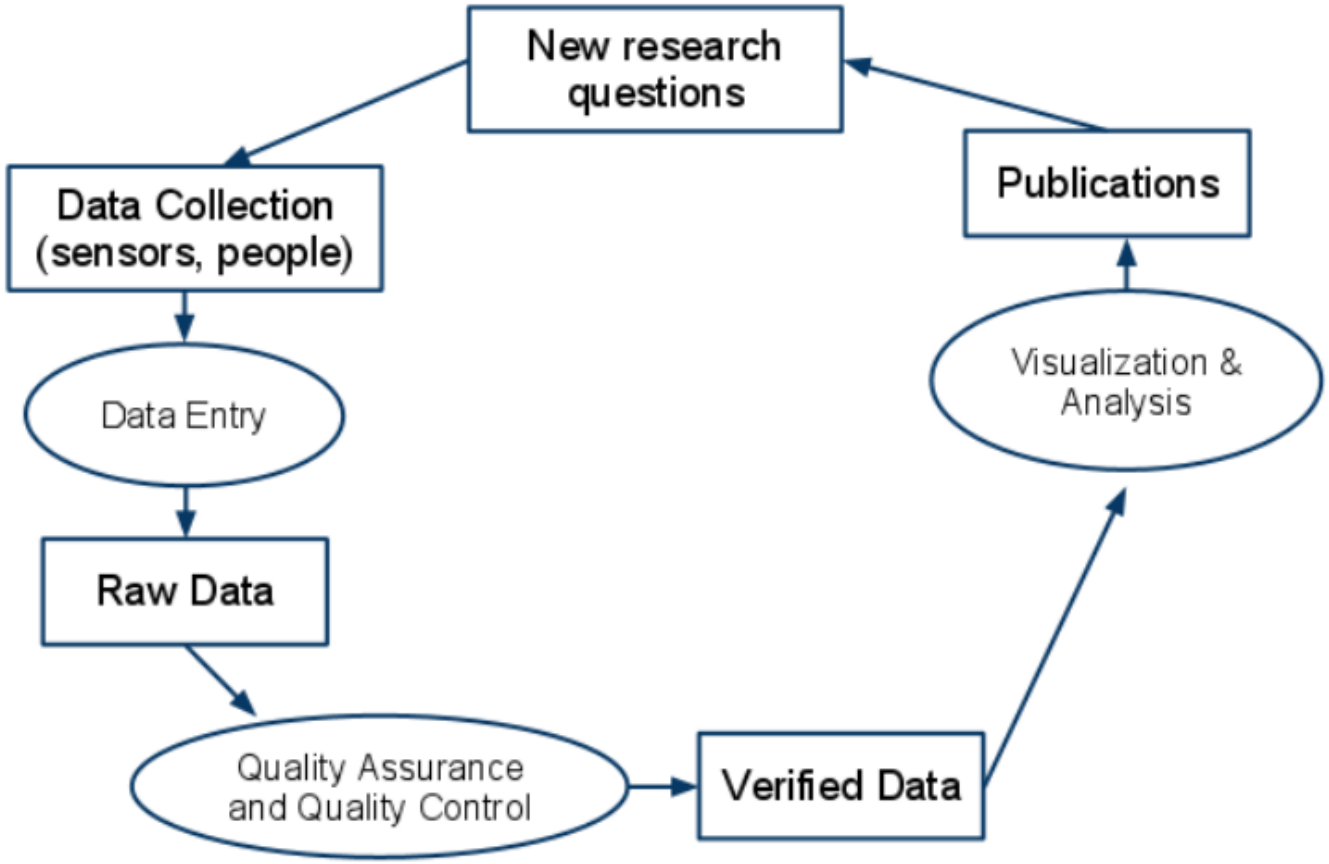
LTER Information Managers Committee

Introduction to LTER Data Management

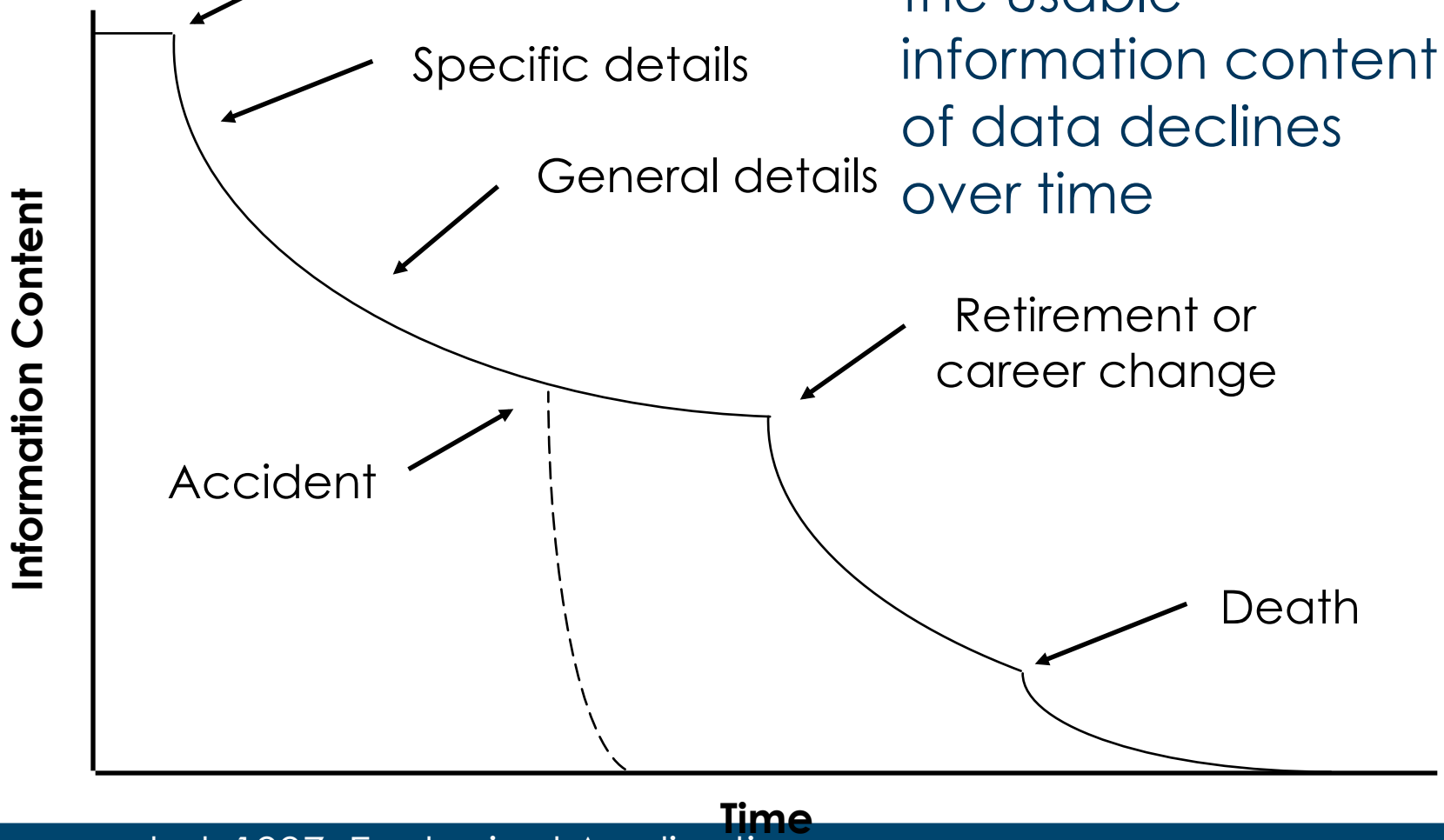
Shamelessly scavenged from presentations by John Porter, Kristin Vanderbilt, Hook et. al, and DataONE

Scientific Use of Data

- The traditional model of using data



Challenge: Understanding Data

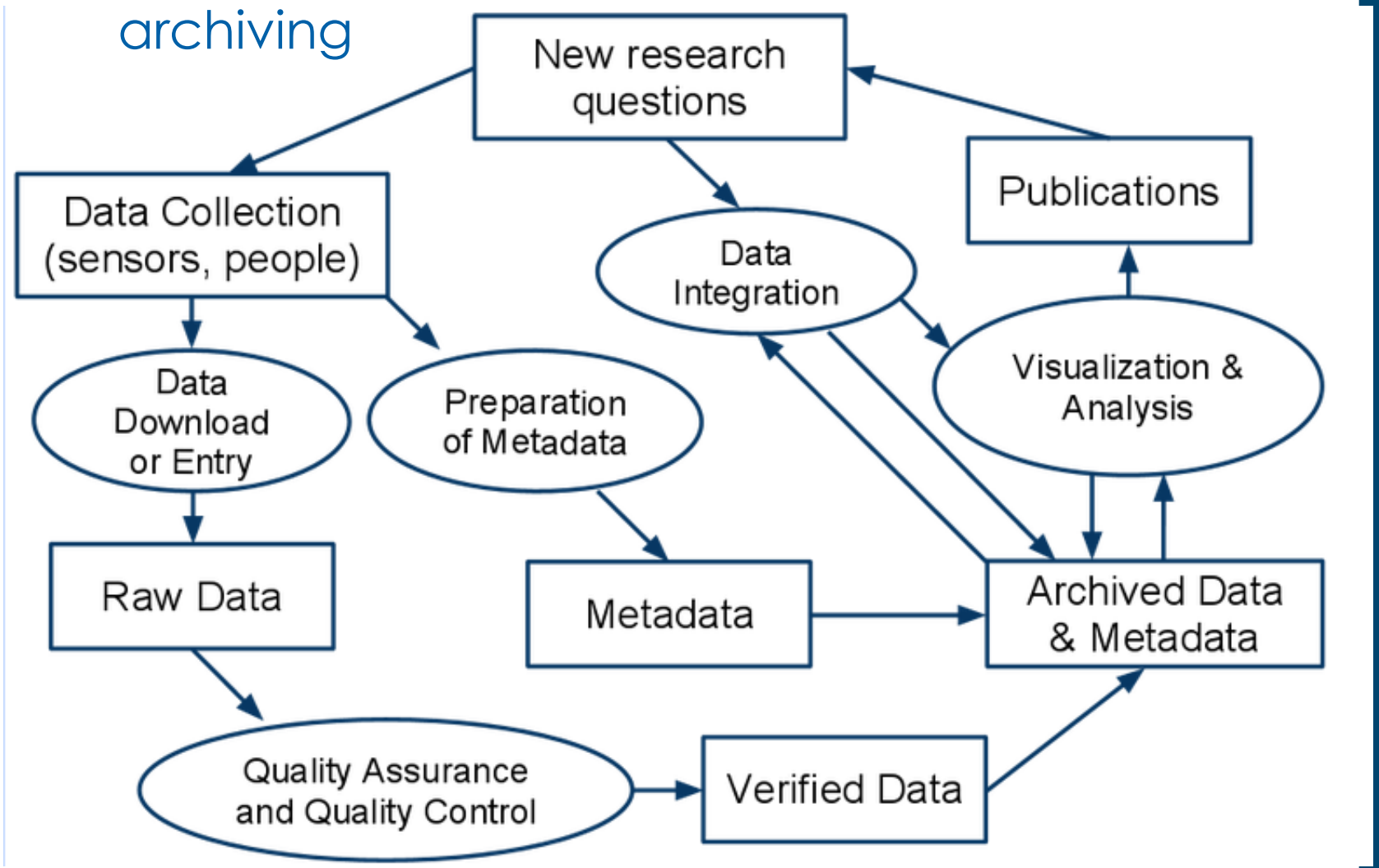


Without Metadata, the usable information content of data declines over time

Michener et al. 1997. Ecological Applications

Scientific Use of Data

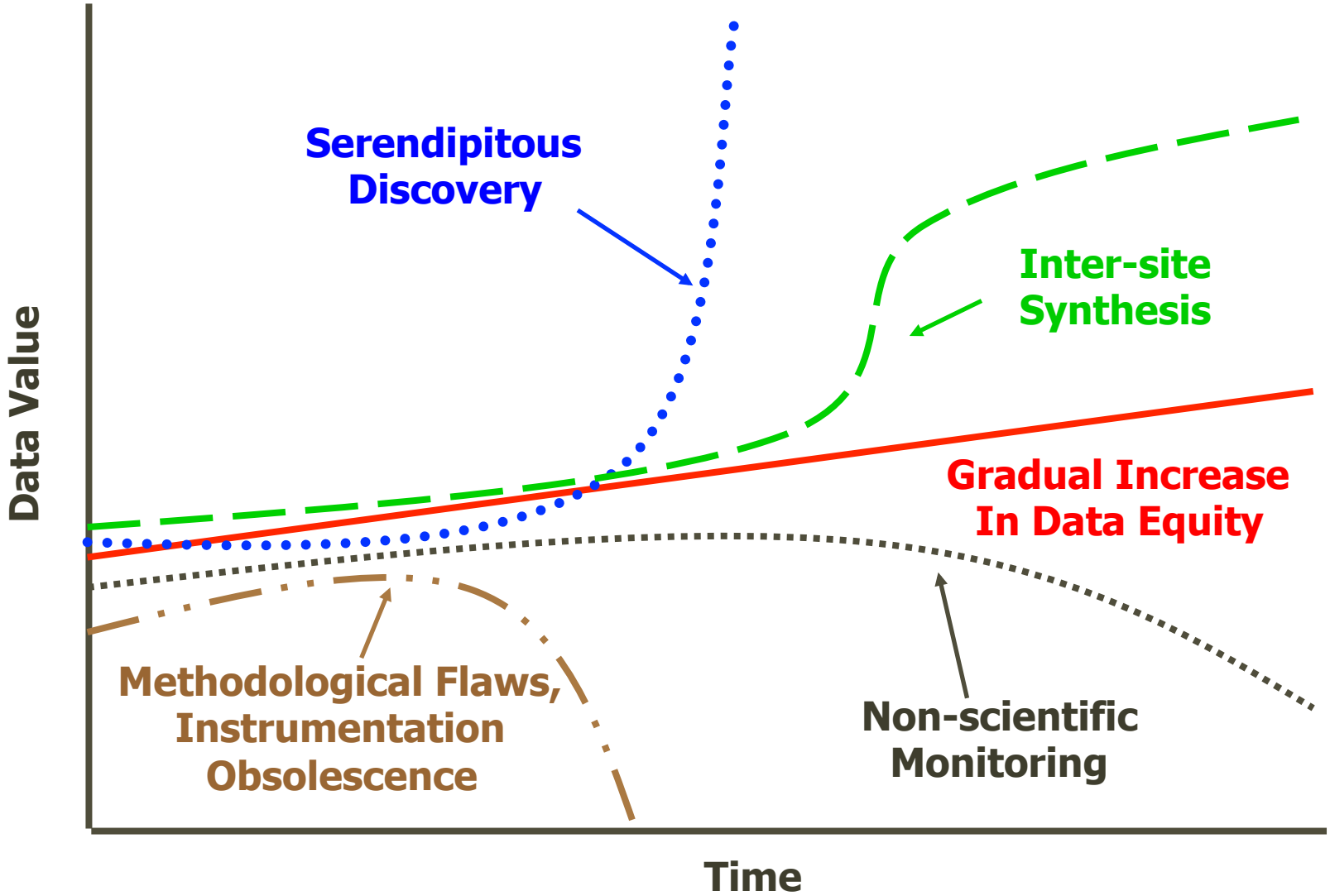
- A new model incorporates sharing and archiving



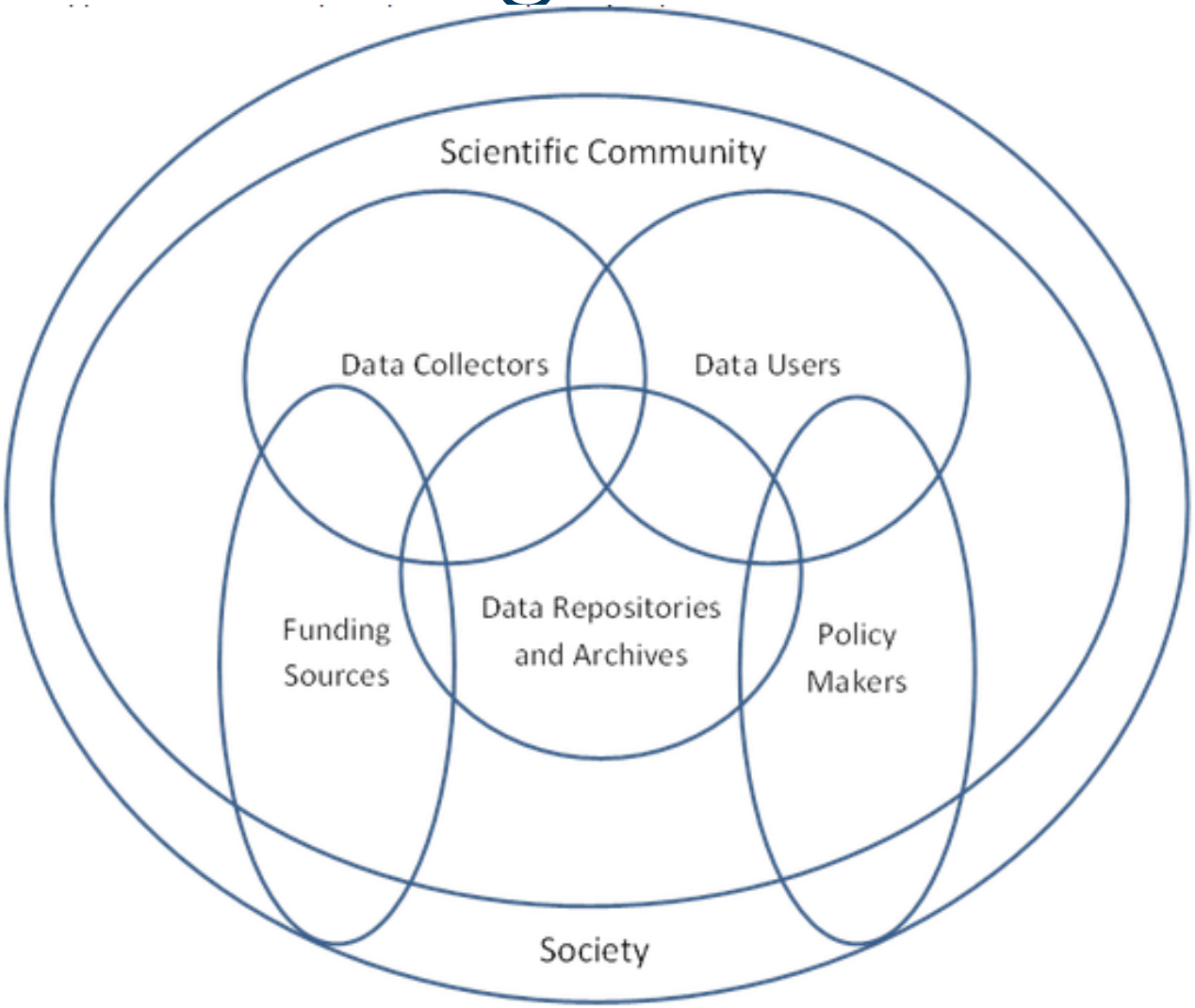
Michiner et. al. 2011, Ecological Informatics

Slide from James Brunt

✓ Increasing value of data over time



Data Sharing and Archiving



Data Sharing

- The LTER Network Data Policy dictates that almost all data should be made available within 2-years
  - exceptions must be justified
- NSF now requires Data Management Plans for LTER
- A better plan increases your chance of funding

Additional Incentives

- NSF now requires Data Management Plans for non-LTER data as well
  - A better plan increases your chance of funding
- Journals are increasingly requiring data submission as a condition of publication for papers (e.g,., evolution, genomics journals)
- Increasingly data is citable
  - Allows you to tally the citations of your data as well as citations of your publications
- Data can even be published: e.g., Ecological Archives publishes "data papers" that are peer-reviewed

Value of Data Sharing

- Open science and new research
- Data longevity
- Data reusability
- Greater exposure to data
- Potential increased citation of source papers (Piwowar, 2007)
- Confirmation of results from publications (Thornton et al., 2005)
- Generation of value added products
- Possibility for future research collaborations
- More value for the sponsor's research investment

- Benefits Pay Off when Data is Well Documented and Well Structured…

But First…Backup Your Data

- Have a working version and at least two backup copies
- Have your data in three separate physical places: **Here, Near,** and **There** *[Brunt's Axiom]*
  - *Computer, a local physically separate backup, and far, far away (e.g., in the Cloud--Dropbox, SugarSync, Mozy, Carbonite)*
  - *Some agencies have restrictions on use of the Cloud. Follow your local guidelines.*

Preparing Environmental Data Sets: 7 Best Practices

- Use Consistent Data Organization
- Use Consistent File Structure
- Use Descriptive File Names
- Assign Descriptive Data Set Titles
- Perform Basic Quality Assurance
- Provide Documentation (JRN LTER Project & Dataset forms)
- Define the Contents of Your Data Files

Organization by Data Type

- If collecting observations of different types of measurements at a site (e.g., leaf area index and above- and belowground biomass), place each type of measurement in a separate data file.
- For each data file, use similar data organization, parameter formats, and common site names, so that users understand the interrelationships between data files.

Organization by Data Type …
(cont'd)

- Data types collected on different time bases (e.g., per hour, per day, per year) are handled more efficiently in separate files.

Use Consistent File Structure

- Use the same structure throughout the file - don't have a different number of columns or re-arrange the columns within the file.
- Use a consistent structure across all data files prepared for a study or project
- Figures and analyses should be reported in companion documents - don't place figures or summary statistics in the data file.

Assign Descriptive File Names

- File names should reflect the contents of the file and include enough information to uniquely identify the data file.
- Clear, descriptive, and unique file names may be important later 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. Avoid using file names such as mydata.dat or 1998.dat.

Assign Descriptive Data Set Titles

- Be aware that your data sets may be accessed many years in the future by people who will be unaware of the details of the project.
- Limit length of the title to 85 characters (spaces included) to be compatible with clearinghouses of ecological and global change data collections.

Define Contents of Your Data Files

In order for others to use your data, they must fully understand the contents of the data set, including the parameter names, units of measure, formats, and definitions of coded values.

Use Consistent Data Organization

- Each row in a file represents a complete record, and the columns represent all the parameters that make up the record. This arrangement is similar to a spreadsheet or matrix.
- In general, keep a set of similar measurements together (e.g., same investigator, methods, time basis, and instruments) in one data file. Please do not break up your data into many small files, e.g., by month or by site.

Data Structure

- The ways researchers typically use data are frequently not compatible with best practices for archiving [the Confetti Toss—one time usage]

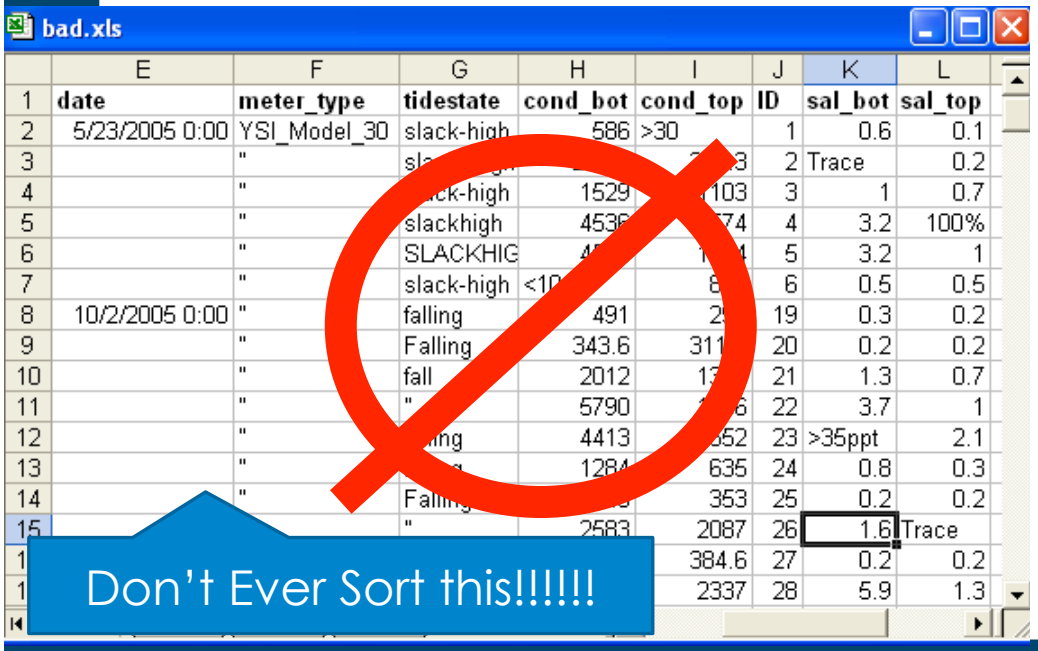
ugly.xls

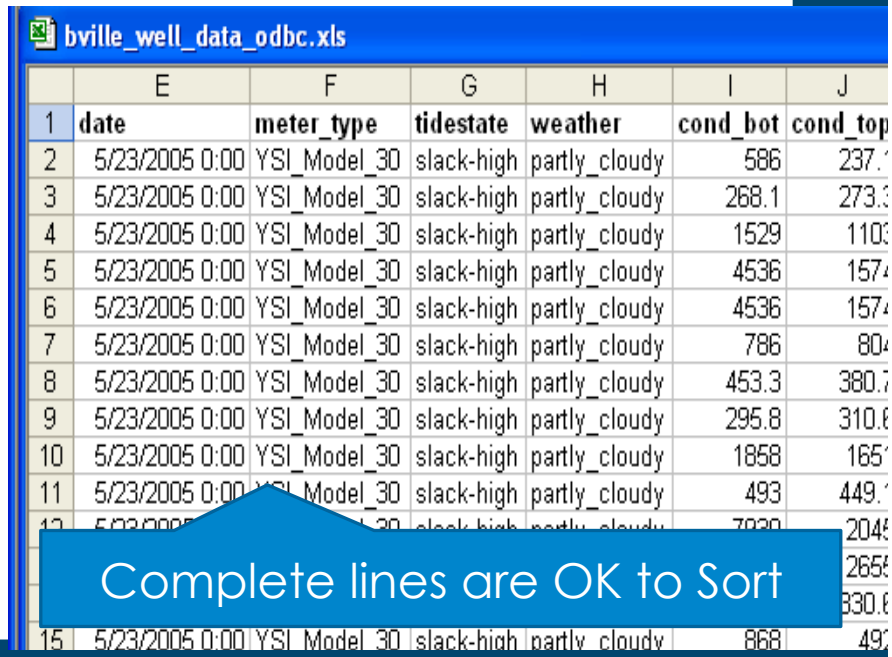
	E	F	G	H	I	J
1	**My Ugly Data Spreadsheet**					
2	Date: 5/23/2005					
3	Meter Type	YSI_Model_30		cond_bot	586	
4	Tide State	slack-high		conductivity_top	<30	
5	Salinity	Bottom	0.6	Top	0.1	
6						
7	Date:	10/2/2005				
8	TIDESTATE	-0.34		MeterType	YSI Model 30	
9	Salinity_bottom	0.3		Conductivity Top	349	
10	Salinity_Top	None		cond_bot	39%	
11						

Sheet1 / Sheet2 / Sheet3

LTER Solutions

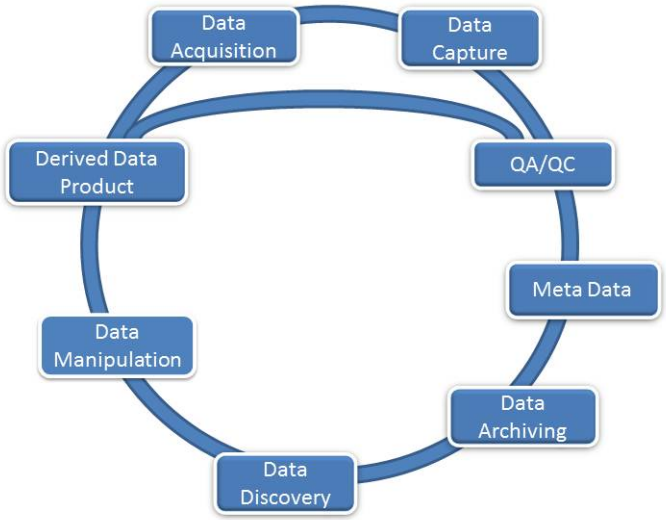
- Data manager can help vet data
- Help communicate best practices to students and investigators
- Use of improved tools that encourage good practices





QA/QC

- Mechanisms designed to prevent introduction of errors in to a dataset



Perform Basic Quality Assurance

- **QA for Tabular Data**
  - Ensure metadata accurately reflects data file names and the actual data file's format, column headers, and content
  - Check that data are in the proper columns
  - Check for & document missing values
  - Check validity of values
  - **Verify** data (from field notebooks, dataloggers, or instruments)

Loss of data quality can occur at many stages:

- At the time of collection
  - Incorrect or inaccurate data are entered into a dataset
    - Misidentifying quad
    - Malfunctioning instrumentation
      - Sensor drift
      - Low batteries
      - Damage

Loss of data quality can occur at many stages:

- During digitization
  - Mistyping code
  - Transposing numbers

Loss of data quality can occur at many stages:

- During documentation
  - Omission: Data or metadata are not recorded
    - Inadequate documentation of experimental design, sampling methods
    - Inadequate documentation of anomalies in the field
    - Forgetting to take measurement in the field

Book3

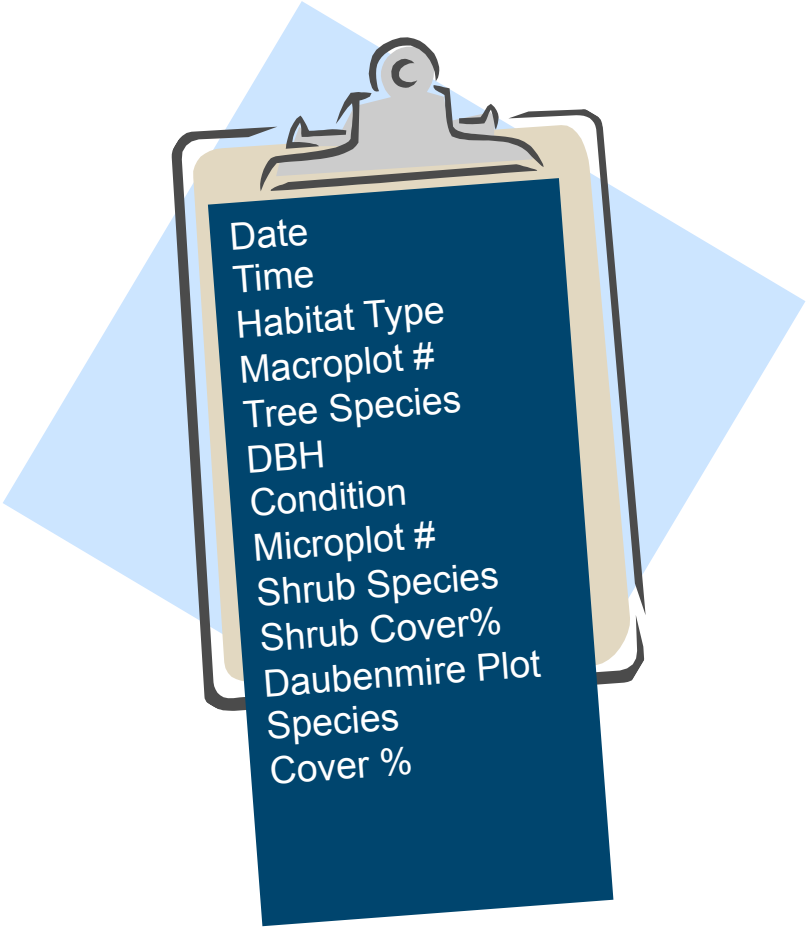
	A	B	C	D	E	F
1	Date	Tree_Number	Species	DBH	Condition	
2	8/6/2012	1	ACRU	56.2	Healthy	
3	8/6/2012	2	THPL	47.1	Disease Evident	
4	8/6/2012	3	THPL	12.1	Forgot to record	
5	8/6/2012	4	TSME	13	Forgot to record	
6						

- Datasheet Design
- Training technicians
- Documentation of procedures
- Data Entry Constraints

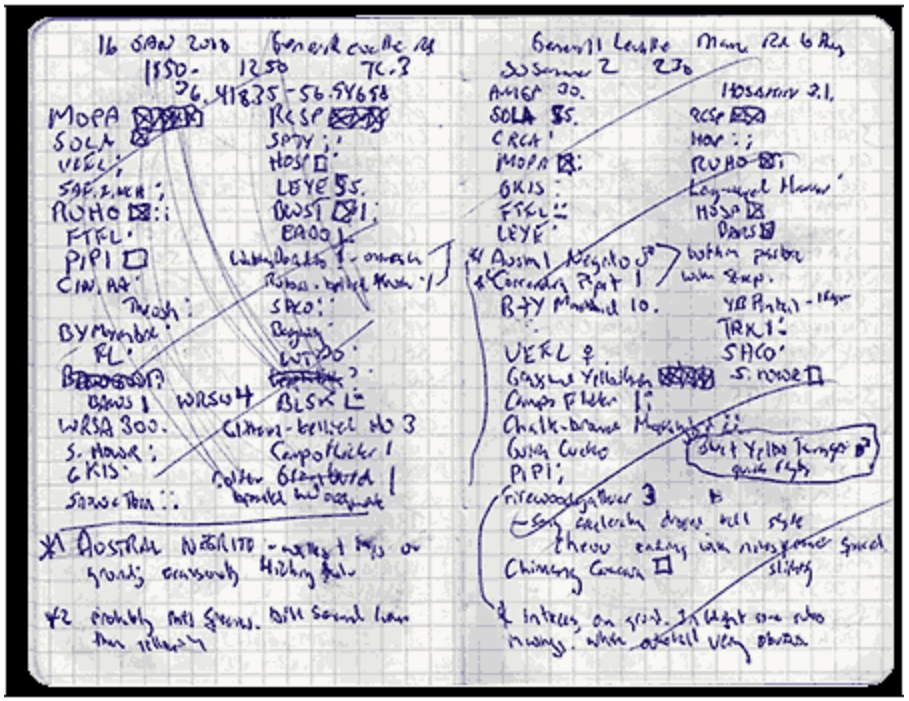
- Graphics
- Statistics

Cost of error correction increases

Datasheets facilitate data collection



Hard to transcribe….



Emulate data file structure…

Type	Study#	StudyName	Year	Qtr	Date	Site	Time	Tower	Height_(cm)	Avg_Bag_Wt_(g)	Bag+Sample_(g)	Sample_Wt_(g)	QWt	Comments
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	944	East	5	4.431	5.341	0.910	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	944	East	10	4.431	4.931	0.500	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	944	East	20	4.431	5.734	1.303	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	944	East	50	4.431	5.339	0.908	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	944	East	100	4.431	5.495	1.064	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	947	Middle	5	4.431	5.190	0.759	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	947	Middle	10	4.431	4.852	0.421	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	947	Middle	20	4.431	5.531	1.100	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	947	Middle	50	4.431	5.622	1.191	G	
BSNE	122	ScrapeSite	2013	1	2/26/2013	Scrape Site	947	Middle	100	4.431	6.058	1.627	G	

Includes descriptive header

Study: BNSE dust collectors - NPP sites, Burn site, Geomet grid data	
Project: Study #288 - BSNE dust collection at NPP, JER Pasture 13, Geomet sites	
Responsible investigator: Greg Okin	
Attribute description:	
Type	= sampling type (core=soil core; BSNE=BSNE dust collector)
Study#	= LTER Study number (122 (SS), 228, 288, 308, ?)
StudyName	= short name for study (ConnPilot, Geomet, LTER-I_TranProbeSoil, NPP, P11aWindErosion, P13Burn, S
Year	= year of field collection (yyyy)
Qtr	= quarterly collection period (1,2,3,4)
Date	= date of field collection (mm/dd/yyyy)
Site	= Scrape Site
Time	= time of field collection (24-hour clock)
Tower	= Tower (East, Middle, West)
Height	= BSNE collector height (5,10,20,50,100 cm above ground when opening oriented to southwest)
Avg_Bag_Wt	= average bag weight (in grams recorded to 3 decimal places)
Bag+Sample	= bag + sample weight (in grams recorded to 3 decimal places)
Sample_Wt	= air dry weight of dust (grams) calculated from (Bag+Sample) minus (Avg_Bag_Wt)
QWt	= data quality flag for Sample Wt
	G = Good; no problems; normal collection
	TL = Not collected; **TOO LITTLE** to collect (less than a "pinch"); Any dust (if any) in trap was emptied
	PC = Partial collection; <u>some</u> sample lost <u>during</u> collection
	...
Comments	= Explain QWt entry or other observations
missing value = .	

Provide Data Set Documentation and Metadata

- Project and Dataset documentation/ metadata forms available online.
- Includes “who, what, where, when, why, and how” of the data.
- Write it for a user 20 years into the future.

Submit and Archive Data with Metadata Early in Your Study

- Submit Project & Dataset documentation within 90 days of initiation of data collection
- First time data submission must be submitted with updated Jornada metadata forms previously submitted.
- First time data submission should be no later than 3 months prior to graduation

Open vs Restricted Access Data

- Open Access Data is made available online.
- Restricted Access Data is not made available to others without the release authorization of the responsible investigator.
  - Data backed up or archived with Jornada less than 2 years after submission or with exemption by LTER Exec Committee.
  - Students have opportunity to publish data following graduation

Summary

- If data are:
  - Well-organized
  - Documented
  - Preserved
  - Accessible
  - Verified as to Accuracy and validity
- Result is:
  - High quality data
  - Easy to share and re-use in science
  - Citation and credibility to the researcher

Acknowledgments:

LTER slides downloaded from http://im.lternet.edu/node/937 from presentations by John Porter, Kristin Vanderbilt, and DataONE.

Downloaded from http://im.lternet.edu/node/948

Les A. Hook, 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.