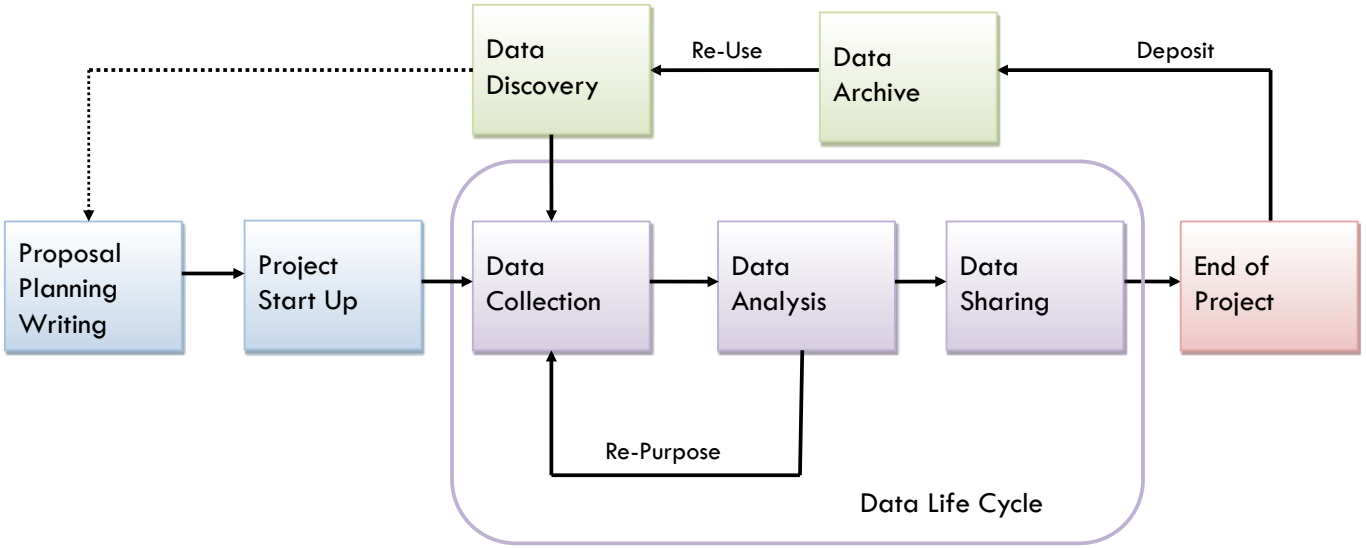
Best Practices for Collecting Data



Bill Corey
Data Consultant
University of Virginia Library
wtc2h@virginia.edu

Andrea Horne Denton
Health Sciences Data Consultant
Claude Moore Health Sciences Library
ash6b@virginia.edu

Goals for the workshop

- Learn about why this is important
- Learn about common problems
- Learn about 7 best practice areas
- Complete hands-on exercises
- Gain peer and expert feedback

Website with Sample Files

Go to:

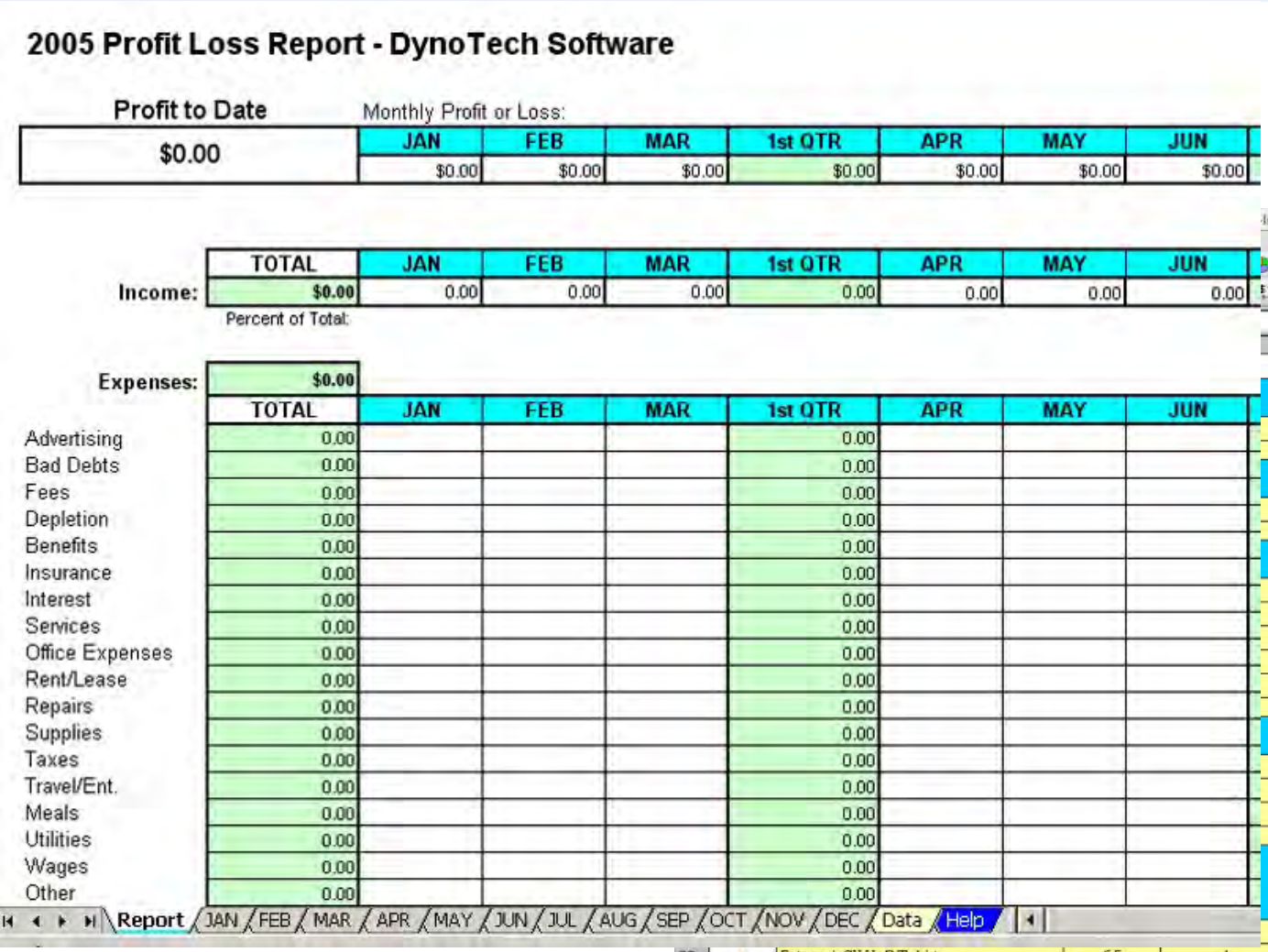
http://dmconsult.library.virginia.edu/best-practices-workshop/

WHY?

Following these Best Practices…….

- Will improve the usability of the data by you or by others
- Your data will be “computer ready”

Spreadsheet Examples

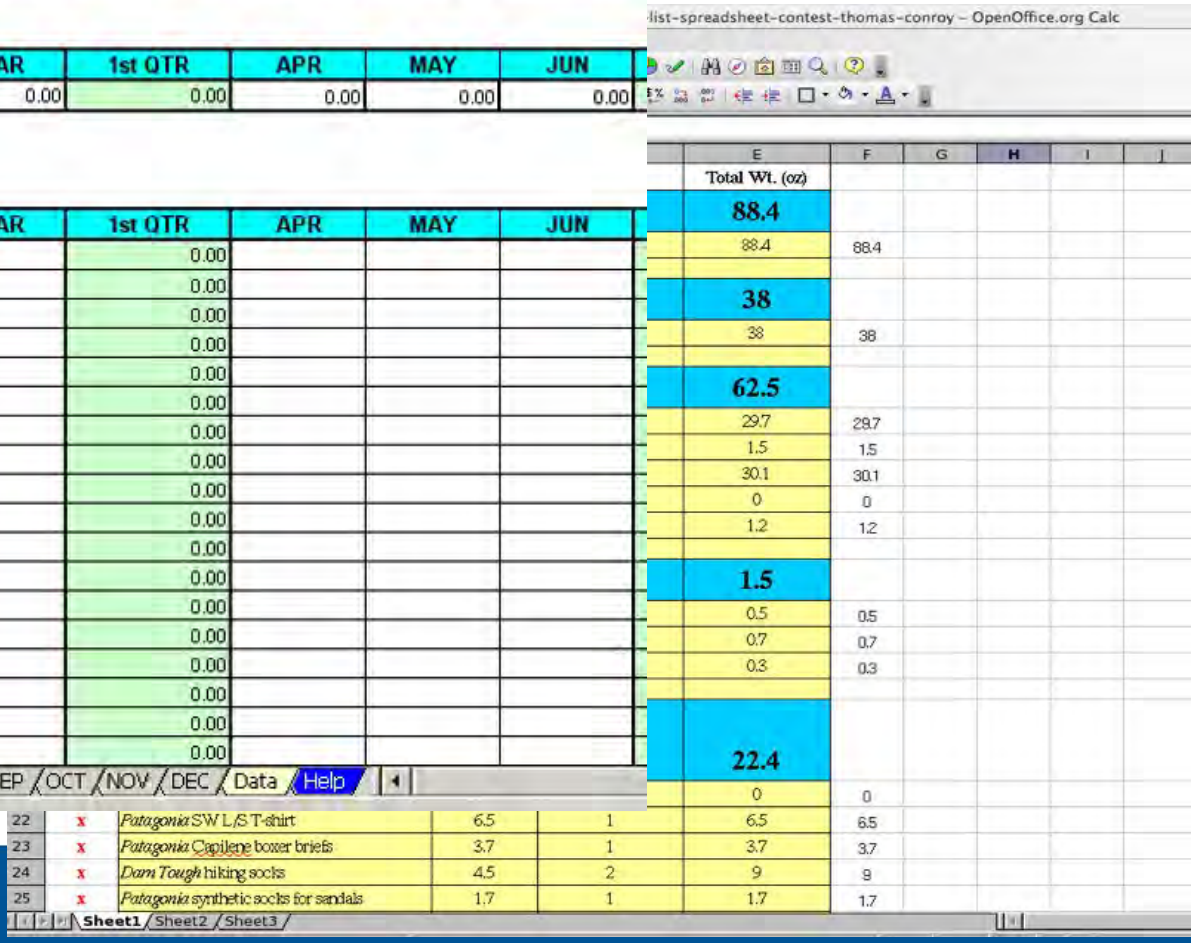


2005 Profit Loss Report - DynoTech Software

Profit to Date	Monthly Profit or Loss:						
$0.00	JAN	FEB	MAR	1st QTR	APR	MAY	JUN
	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00	$0.00

	TOTAL	JAN	FEB	MAR	1st QTR	APR	MAY	JUN
Income:	$0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Percent of Total:							
Expenses:	$0.00							
	TOTAL	JAN	FEB	MAR	1st QTR	APR	MAY	JUN
Advertising	0.00				0.00			
Bad Debts	0.00				0.00			
Fees	0.00				0.00			
Depletion	0.00				0.00			
Benefits	0.00				0.00			
Insurance	0.00				0.00			
Interest	0.00				0.00			
Services	0.00				0.00			
Office Expenses	0.00				0.00			
Rent/Lease	0.00				0.00			
Repairs	0.00				0.00			
Supplies	0.00				0.00			
Taxes	0.00				0.00			
Travel/Ent.	0.00				0.00			
Meals	0.00				0.00			
Utilities	0.00				0.00			
Wages	0.00				0.00			
Other	0.00				0.00			

Report | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC | Data | Help



list-spreadsheet-contest-thomas-conroy – OpenOffice.org Calc

	E	F	G	H	I	J
	Total Wt. (oz)					
	88.4					
	88.4	88.4				
	38					
	38	38				
	62.5					
	29.7	29.7				
	1.5	1.5				
	30.1	30.1				
	0	0				
	1.2	1.2				
	1.5					
	0.5	0.5				
	0.7	0.7				
	0.3	0.3				
	22.4					
	0	0				
22	x	*Patagonia* SW L/S T-shirt	6.5	1	6.5	6.5
23	x	*Patagonia* Capilene boxer briefs	3.7	1	3.7	3.7
24	x	*Darn Tough* hiking socks	4.5	2	9	9
25	x	*Patagonia* synthetic socks for sandals	1.7	1	1.7	1.7

Sheet1 | Sheet2 | Sheet3

Spreadsheet Problems?

messydata.xlsx

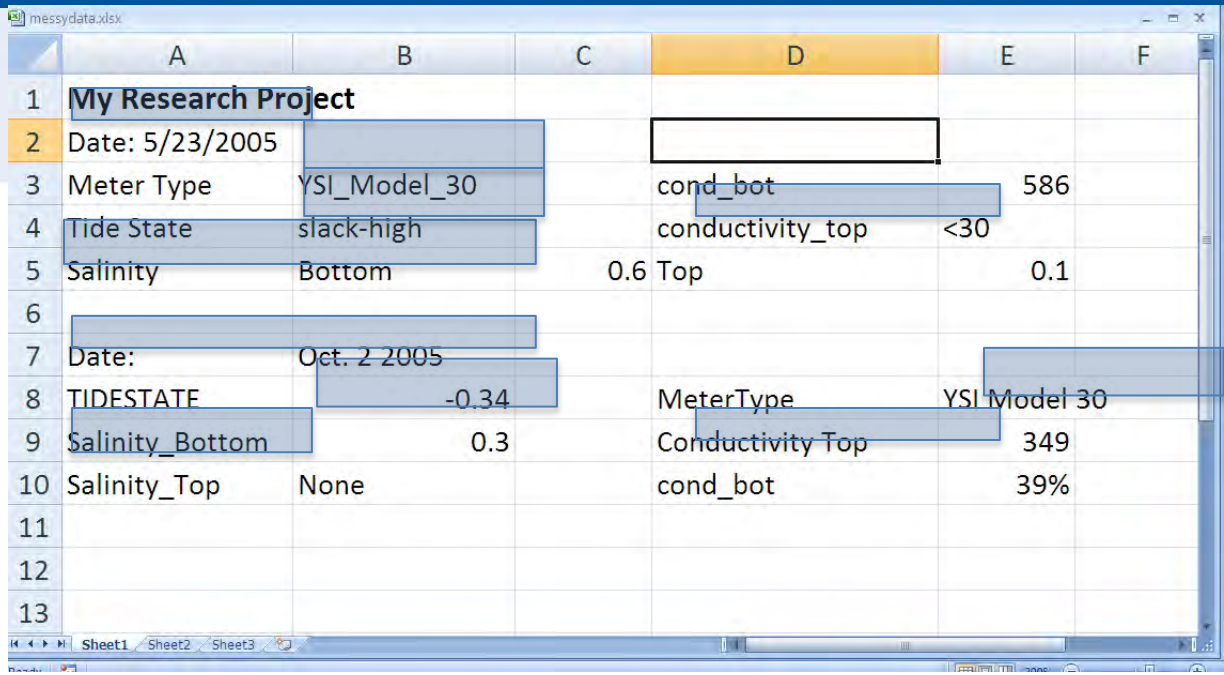
	A	B	C	D	E
1	**My Research Project**				
2	Date: 5/23/2005				
3	Meter Type	YSI_Model_30		cond_bot	586
4	Tide State	slack-high		conductivity_top	<30
5	Salinity	Bottom	0.6	Top	0.1
6					
7	Date:	Oct. 2 2005			
8	TIDESTATE	-0.34		MeterType	YSI Model 30
9	Salinity_Bottom	0.3		Conductivity Top	349
10	Salinity_Top	None		cond_bot	39%
11					
12					
13					

Sheet1 Sheet2 Sheet3

Pause for Exercise

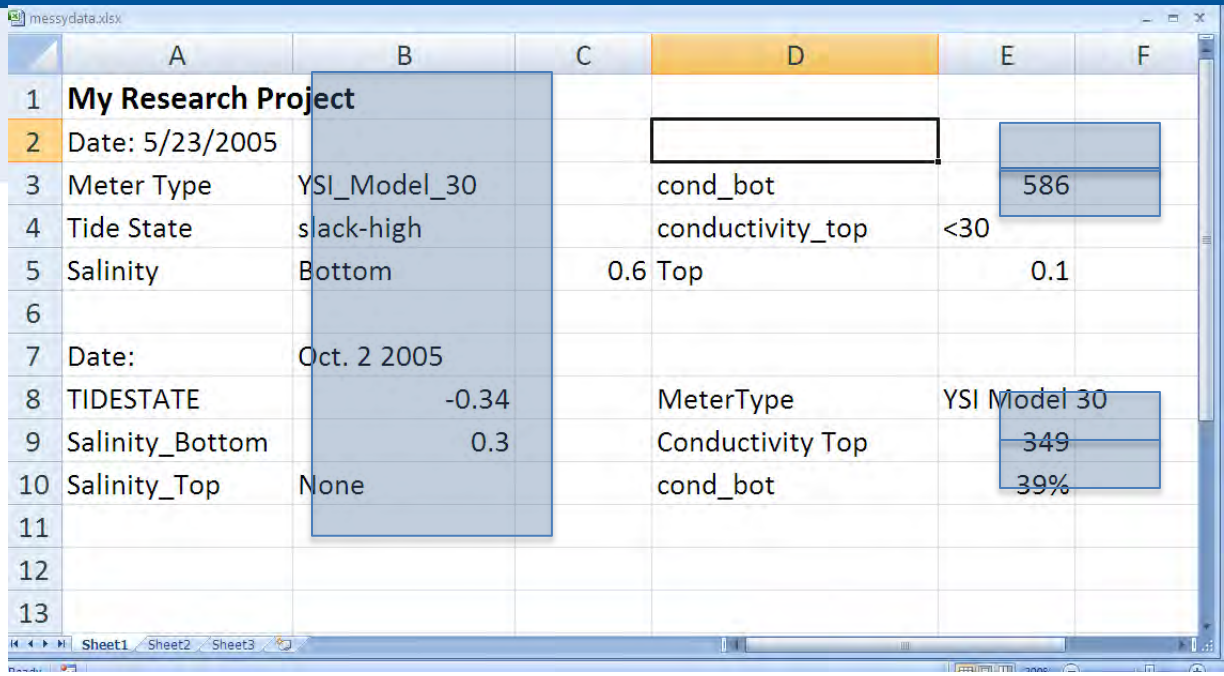
Problems

- Dates are not stored consistently



- Values are labeled inconsistently
- Data coding is inconsistent
- Order of values are different

Problems



- Confusion between numbers and text
- Different types of data are stored in the same columns
- The spreadsheet loses interpretability if it is sorted

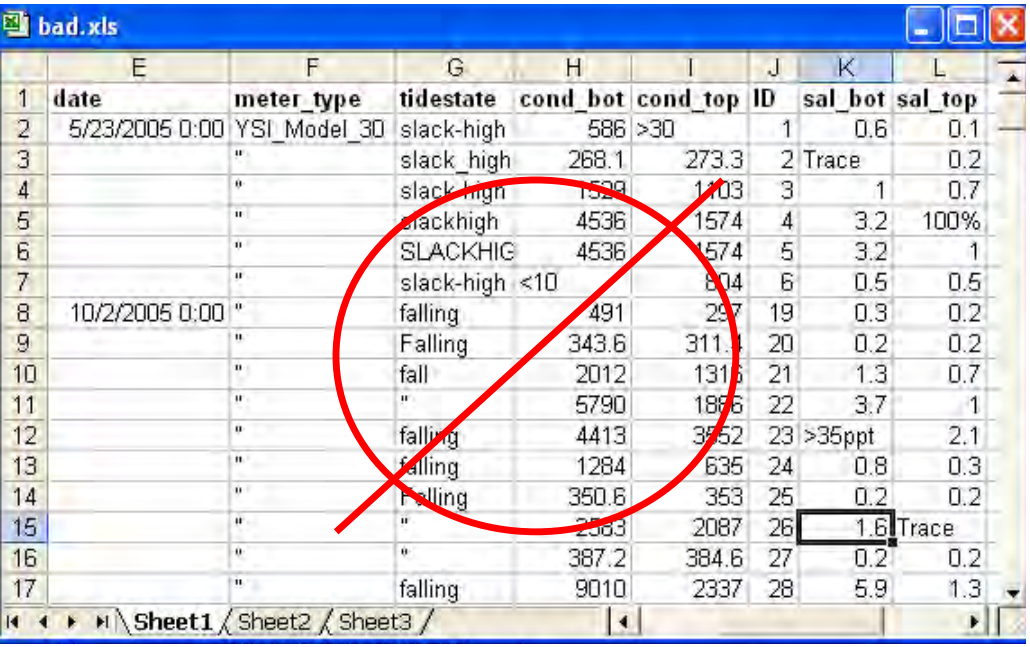
Possible Solution

	A	B	C	D	E	F	G
1	Date	Meter Type	TideState	Salinity_Bottom	Salinity_Top	Conductivity_Top	Conductivity_Bottom
2	5/23/05	YSI_Model_30	0.1	0.6	0.1	28	586
3	10/2/05	YSI_Model_30	-0.34	0.3	0	349	200
4							
5							

Next Exercise

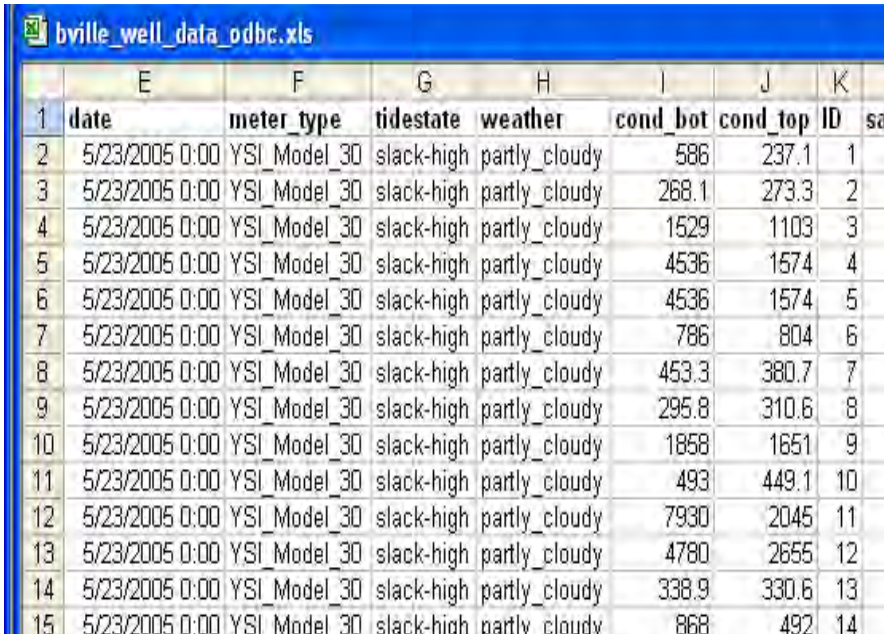
Best Practices Data Organization

- Lines or rows of data should be complete
  - Designed to be machine readable, not human readable (sort)


bad.xls

	E	F	G	H	I	J	K	L
1	date	meter_type	tidestate	cond_bot	cond_top	ID	sal_bot	sal_top
2	5/23/2005 0:00	YSI_Model_30	slack-high	586	>30	1	0.6	0.1
3		"	slack_high	268.1	273.3	2	Trace	0.2
4		"	slack high	1529	1103	3	1	0.7
5		"	slackhigh	4536	1574	4	3.2	100%
6		"	SLACKHIG	4536	1574	5	3.2	1
7		"	slack-high	<10	804	6	0.5	0.5
8	10/2/2005 0:00	"	falling	491	297	19	0.3	0.2
9		"	Falling	343.6	311.4	20	0.2	0.2
10		"	fall	2012	1313	21	1.3	0.7
11		"	"	5790	1866	22	3.7	1
12		"	falling	4413	3652	23	>35ppt	2.1
13		"	falling	1284	635	24	0.8	0.3
14		"	Falling	350.6	353	25	0.2	0.2
15		"	"	2583	2087	26	1.6	Trace
16		"	"	387.2	384.6	27	0.2	0.2
17		"	falling	9010	2337	28	5.9	1.3

Sheet1 / Sheet2 / Sheet3


bville_well_data_odbc.xls

	E	F	G	H	I	J	K
1	date	meter_type	tidestate	weather	cond_bot	cond_top	ID
2	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	586	237.1	1
3	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	268.1	273.3	2
4	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	1529	1103	3
5	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	4536	1574	4
6	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	4536	1574	5
7	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	786	804	6
8	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	453.3	380.7	7
9	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	295.8	310.6	8
10	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	1858	1651	9
11	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	493	449.1	10
12	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	7930	2045	11
13	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	4780	2655	12
14	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	338.9	330.6	13
15	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	868	492	14

Possible Solution

	A	B	C	D	E	F	G	H	I
1	date	meter_type	tidestate	cond_bot	cond_top	ID	sal_bot	sal_top	
2	2/10/08	YSI_Model_30	slack-high	596	33	1	0.6	0.1	
3	2/10/08	YSI_Model_30	slack-high	268.1	273.3	2	0.1	0.2	
4	2/10/08	YSI_Model_30	slack-high	1529	1102	3	1	0.07	
5	2/10/08	YSI_Model_30	slack-high	4534	1574	4	3.2	1	
6	2/10/08	YSI_Model_30	slack-high	4534	1543	5	3.2	1	
7	2/10/08	YSI_Model_30	slack-high	9	804	6	0.5	0.5	
8	4/23/08	YSI_Model_30	falling	491	297	19	0.3	0.2	
9	4/23/08	YSI_Model_30	falling	343.6	311.3	20	0.3	0.2	
10	4/23/08	YSI_Model_30	falling	2012	1316	21	1.3	0.7	
11	4/23/08	YSI_Model_30	falling	5790	1866	22	3.7	1	
12	4/23/08	YSI_Model_30	falling	4413	3552	23	3.5	2.1	
13	4/23/08	YSI_Model_30	falling	1284	635	24	0.8	0.3	
14	4/23/08	YSI_Model_30	falling	350.5	353	25	1.6	0.1	
15									

Best Practices Data Organization

date	meter_type	tidestate	cond_bot	cond_top	ID	sal_bot	sal_top
2/10/2008	YSI_Model_30	slack-high	596	33	1	0.6	0.1
2/10/2008	YSI_Model_30	slack-high	268.1	273.3	2	0.1	0.2

- Include a Header Line 1st line (or record)
- Label each Column with a short but descriptive name
  - Names should be unique
  - Use letters, numbers, or "_" (underscore)
  - Do not include blank spaces or symbols (+ - & ^ *)

Best Practices Data Organization

date	meter_type	tidestate	cond_bot	cond_top	ID	sal_bot	sal_top
2/10/2008	YSI_Model_30	slack-high	596	33	1	0.6	0.1
2/10/2008	YSI_Model_30	slack-high	268.1	273.3	2	0.1	0.2

- Columns of data should be consistent
  - Use the same naming convention for text data
- Columns should include only a single kind of data
  - Text or “string” data
  - Integer numbers
  - Floating point or real numbers

Use Standardized Formats

- ISO 8601 Standard for Date and Time
  - YYYYMMDDThh:mmss.sTZD

    20091013T09:1234.9Z

    20091013T09:1234.9+05:00
- Spatial Coordinates for Latitute/Longitude
  - +/- DD.DDDDD

    -78.476 (longitude)

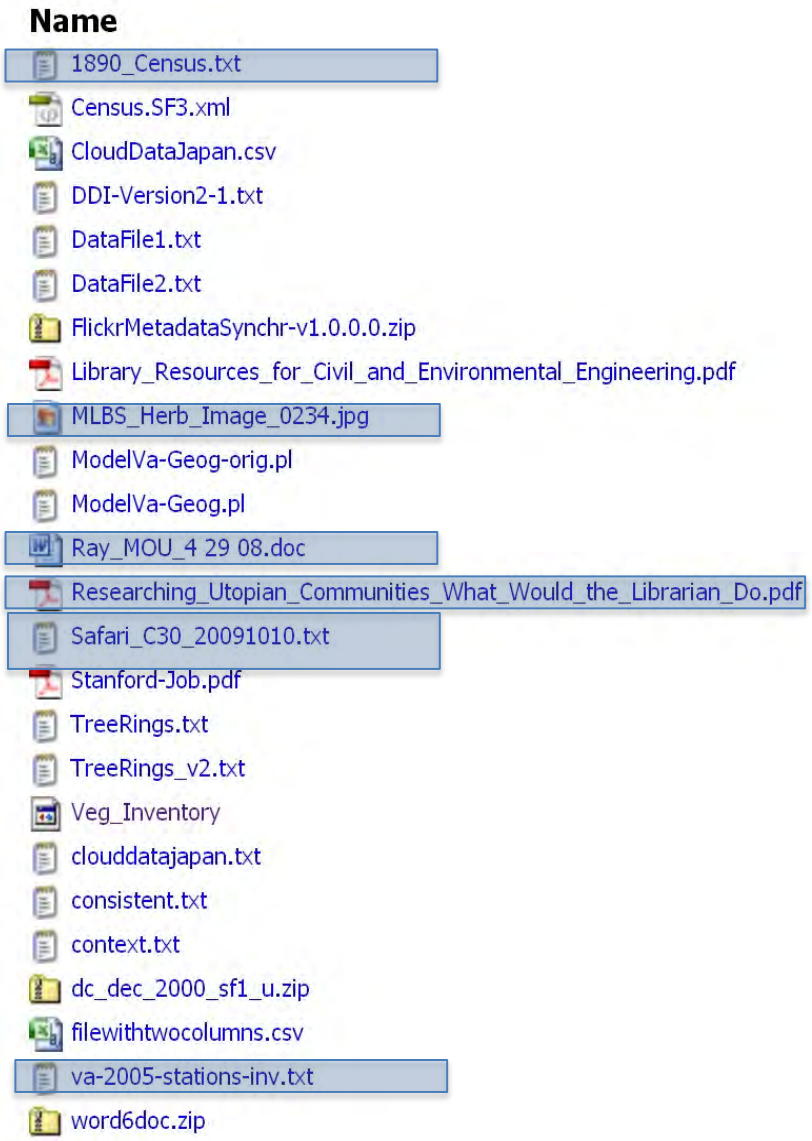
    +38.029 (latitude)

File Names


Name
1890_Census.txt
Census.SF3.xml
CloudDataJapan.csv
DDI-Version2-1.txt
DataFile1.txt
DataFile2.txt
FlickrMetadataSynchr-v1.0.0.0.zip
Library_Resources_for_Civil_and_Environmental_Engineering.pdf
MLBS_Herb_Image_0234.jpg
ModelVa-Geog-orig.pl
ModelVa-Geog.pl
Ray_MOU_4 29 08.doc
Researching_Utopian_Communities_What_Would_the_Librarian_Do.pdf
Safari_C30_20091010.txt
Stanford-Job.pdf
TreeRings.txt
TreeRings_v2.txt
Veg_Inventory
clouddatajapan.txt
consistent.txt
context.txt
dc_dec_2000_sf1_u.zip
filewithtwocolumns.csv
va-2005-stations-inv.txt
word6doc.zip

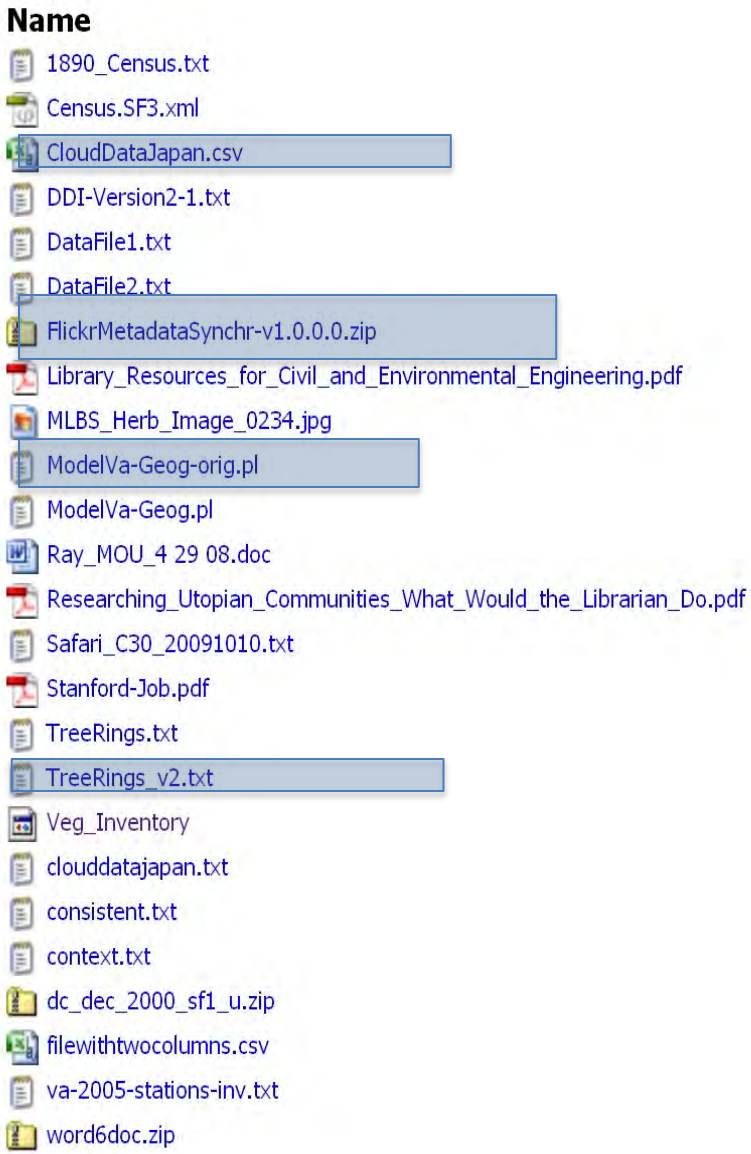
File Names

- Use descriptive names
- Not too long
- Don’t use spaces
- Try to include time, place & theme
- May use “-” or “_”



File Names

- String words together with Caps (VegBiodiv_2007)
- Think about using version numbers
- Don't change default extensions (txt, jpg, csv,…)



Organize Files Logically

- Make sure your file system is logical and efficient

Biodiv_H20_heatExp_2005_2008.csv
Biodiv_H20_predatorExp_2001_2003.csv
…

Biodiv_H20_planktonCount_start2001_active.csv
Biodiv_H20_chla_profiles_2003.csv
…

Quality Assurance / Control

- QA: Manually check 5 – 10% of data records
- QA: Check for out-of-range values (plotting)
- QA: Map Location Data
- QC: Use a data entry program
  - Program to catch typing errors
  - Program pull-down menu option
- QC: Double entry keying

Preserve Information

- Keep Original (Raw) File
  - Uncorrected copy, make “read-only”
- Use scripted code to transform and correct data
- Save as a new file

Raw Data File

TAX	COUNT	TEMPC
C	3.97887358	12.3
C	10.8823893	12.8
M	21.7647785	14.2
N	61.6668725	12.9
F	0.97261354	12.7
M	0.53051648	12.1
F	0	11.9
F	43.5295571	13.1

Processing Script (R)

```
–### Giles_zoop_temp_regress_4jun08.r

–### Load data

–Giles<-
read.csv("Giles_zoopCount_Diel_2001_2003.csv")

–### Look at the data

–Giles

–plot(COUNT~ TEMPC, data=Giles)

–### Log Transform the independent variable (x+1)

–Giles$Lcount<-log(Giles$COUNT+1)

–### Plot the log-transformed y against x

–plot(Lcount ~ TEMPC, data=Giles)
```

Preserving: Scripted Notes

- Use a scripted language to process data
  -  – R Statistical package (free, powerful)
  -  – SAS
  -  – MATLAB
- Processing scripts records processing
  - Steps are recorded in textual format
  - Can be easily revised and re-executed
  - Easy to document
- GUI-based analysis may be easier, but harder to reproduce

Define Contents of Data Files

- Create a Project Document File (Lab Notebook)
- Details such as:
  - Names of data & analysis files associated with study
  - Definitions for data and codes (include missing value codes, names) example
  - Units of measure (accuracy and precision)
  - Standards or instrument calibrations

Next Exercise

- Create a Data Dictionary (Document) for the file “sortdata-good”
- Template

FieldName	Definition	Values	Type	Notes
Deme	Log number		char-num	
Date	Current Date		date (YYYYMMDD)	Could this be automated?
StartTime	Start time of visit	Hours 0 - 23	time(hh:mm:ss)	This needs to be the same for ALL "visitors" on the same "visit", as the VisitID is generated from this field
EndTime	End time of visit	Hours 0 - 23	time(hh:mm:ss)	
Observer	initials for all persons entering data at same visit	initials separated by "+" (plus sign)	char	
RecorderID	initials for person entering data		char	

Possible Solution

FieldName	Definition	Values	Type	Notes
ID	System generated for each new record	ID	integer	
date	Date Water Sampled		date (YYYYMMDD)	
meter_type	Meter used for salinity and conductivity measurements see: http://www.fishersci.com/ecomm/servlet/fsproductdetail_10652_642075__-1_0	YSI_Model_30	string	Only meter used in my experiment
tidestate	Height of tide	slack-high (just below high stage - rising) falling (high stage retreating) high (at high tide) low (at low tide)	string	
cond_bot	Conductivity of the bottom of the water level	from 0 to 200mS/cm, with ±0.5% full-scale accuracy	float	
cond_top	Conductivity of the top of the water level	from 0 to 200mS/cm, with ±0.5% full-scale accuracy	float	
sal_bot	Salinity of the bottom of the water level	from 0 to 80ppt with accuracy of ±2% or ±0.1ppt	float	
sal_top	Salinity of the top of the water level	from 0 to 80ppt with accuracy of ±2% or ±0.1ppt	float	

Data Dictionary Example

Column	Description	Units/Format
SITE	k=Kataba forest, p=Pandamatenga, m=Near Maun, e=HOORC/MPG Maun tower, o=Okwa river crossing, t=Tshane, skukuza=Skukuza Flux Tower	text
SPECIES	Scientific name up to 25 characters	text
DATE	Date of measurement	yyyymmdd
BA	Woody plant basal area	m2/ha
SEBA	Standard error of BA	m2/ha
DENSITY	Woody plant density (number of trees per hectare)	number/ha
SEDEN	Standard error of DENSITY (n=42 for KT, n=49 for Skukuza)	number/ha
STEMS	Number of stems per hectare (/ha)	number/ha
HEIGHT	Basal area-weighted average height	m2/ha
WOOD	Aboveground woody plant wood dry biomass	kg/ha
LEAF	Aboveground woody plant leaf dry biomass	kg/ha
LAI	Leaf Area Index calculated by allometry	m2/m2

File Format Sustainability

Types	Examples
Text	ASCII, Word, PDF
Numerical	ASCII, SPSS, STATA, Excel, Access, MySQL
Multimedia	Jpeg, tiff, mpeg, quicktime
Models	3D, statistical
Software	Java, C, Fortran
Domain-specific	FITS in astronomy, CIF in chemistry
Instrument-specific	Olympus Confocal Microscope Data Format

Choosing File Formats

- Accessible Data (in the future)
  - Non-proprietary (software formats)
  - Open, documented standard
  - Common, used by the research community
  - Standard representation (ASCII, Unicode)
  - Unencrypted & Uncompressed

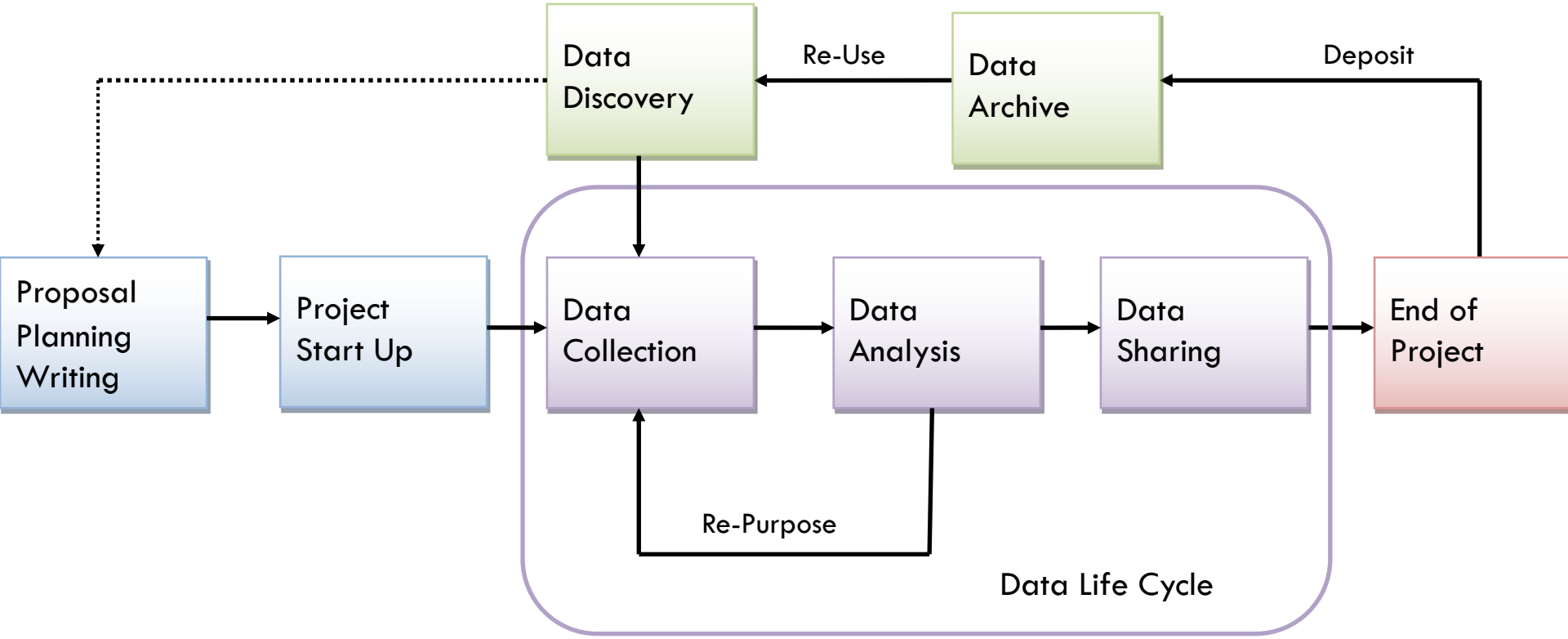
Best Practices Creating Data

1. Use Consistent Data Organization
2. Use Standardized Formats
3. Assign Descriptive File Names
4. Perform Basic Quality Assurance / Quality Control
5. Preserve Information - Use Scripted Languages
6. Define Contents of Data Files; Create Metadata
7. Use Consistent, Stable and Open File Formats

Why Manage Data?

- Saves time
- Others can understand your data
- Makes sharing data easier
  - Increases the visibility of your research
  - Facilitates new discoveries
  - Reduces costs by avoiding duplication
  - Required by funding agencies

Research Life Cycle



Managing Data in the Data Life Cycle

- Choosing file formats
- File naming conventions
- Document and metadata
- Access control & security
- Backup & storage

Data Security & Access Control

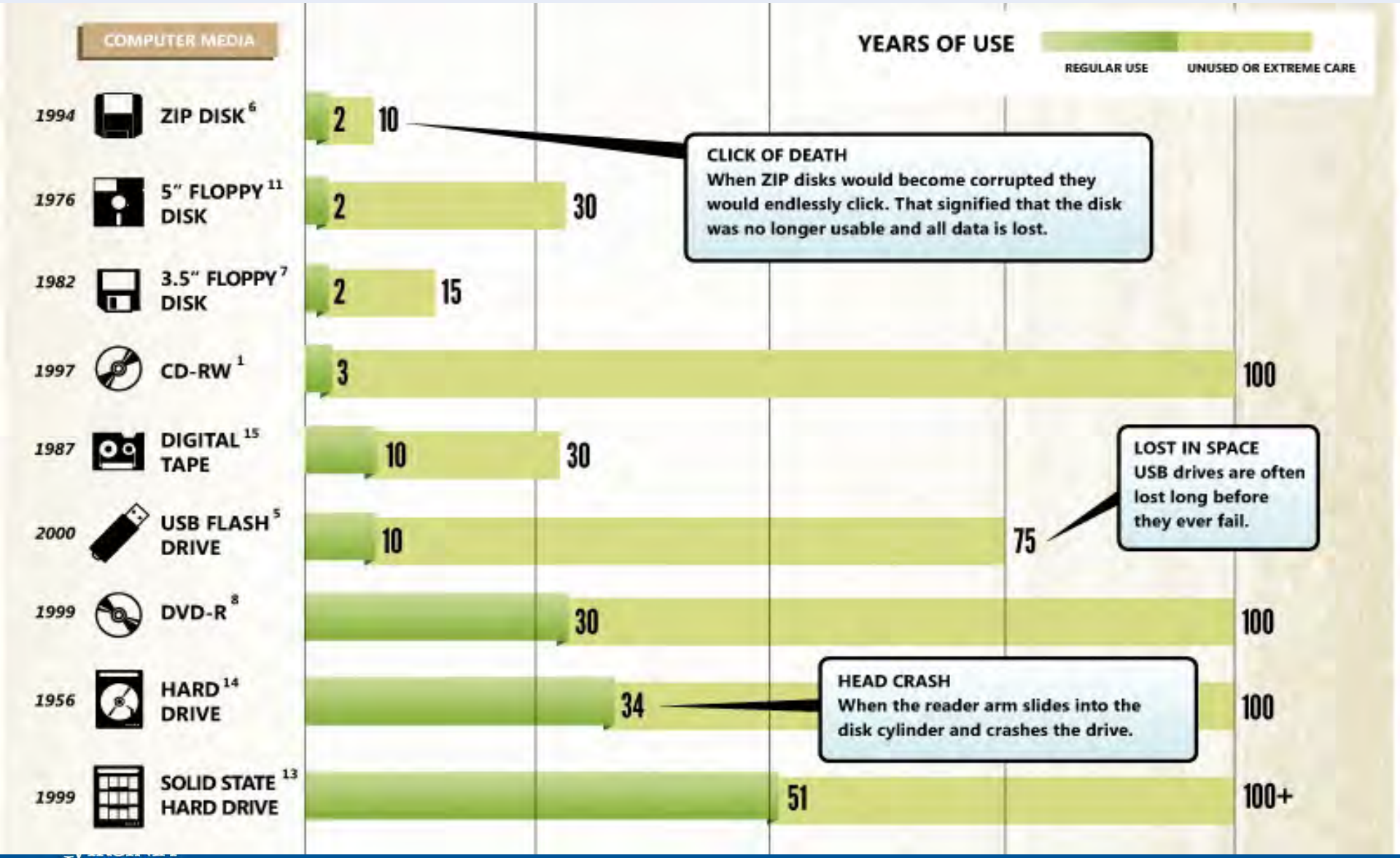
- Network security
  - keep confidential or sensitive data off internet servers or computers on connected to the internet
- Physical security
  - Access to buildings and rooms
- Computer Systems & Files
  - Use passwords on files/system
  - Virus protection

Backup Your Data

- Reduce the risk of damage or loss
- Use multiple locations (here, near, far)
- Create a backup schedule
- Use reliable backup medium
- Test your backup system (i.e., test file recovery)

Storage & Backup

Sustainable Storage



Lifespan of Storage Media: http://www.crashplan.com/medialifespan/

Best Practices Bibliography

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some simple guidelines for effective data management. *Bulletin of the Ecological Society of America, 90*(2), 205-214.

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to social science data preparation and archiving: Best practices throughout the data cycle* (5th ed.). Ann Arbor, MI. Retrieved 05/31/2012, from http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf.

Graham, A., McNeill, K., Stout, A., & Sweeney, L. (2010). *Data Management and Publishing.* Retrieved 05/31/2012, from http://libraries.mit.edu/guides/subjects/data-management/.

Best Practices Bibliography (Cont.)

Van den Eynden, V., Corti, L., Woollard, M. & Bishop, L. (2011). *Managing and Sharing Data: A Best Practice Guide for Researchers* (3rd ed.). Retrieved 05/31/2012, from http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Hook, L. A., Santhana Vannan, S.K., Beaty, T. W., Cook, R. B. and Wilson, B.E. (2010). *Best Practices for Preparing Environmental Data Sets to Share and Archive*. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. http://dx.doi.org/10.3334/ORNLDAAC/BestPractices-2010.

Mailing List Subscription

- Please check the box on our sign-in sheet to receive occasional emails to keep up with our services, training, and news.

- Please encourage others to subscribe: http://eepurl.com/CJwYT

More Research Data Services in the Library

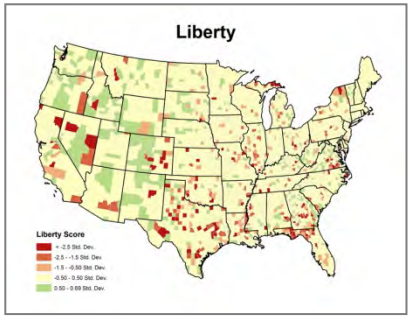
Offering expert data assistance at every stage of the research process.

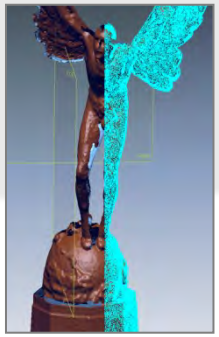
PLANNING

Need a data management plan?

We can assist you with developing a data management plan that meets increasingly stringent criteria from funding agencies, including:

- Implementation of procedures, tools, and workflows for managing data sets
- Designing a strong study that yields reliable statistics





FINDING & COLLECTING

Need help finding data or collecting your own?

We have thousands of sources with the data you seek and experts who will help you:

- Locate, evaluate and format data
- Design metadata and data documentation protocols for new data collection
- Capture data with the appropriate technology tools for your needs

researchdataservices@virginia.edu

SHARING

Ready to share or archive your data?

We can consult with you on strategies to help others discover or access your research by:

- Adhering to data sharing policies and norms
- Selecting a data-sharing repository
- Making your data easier to discover and link

ANALYZING

Want help uncovering unique and compelling insights?

Get expert assistance from statistical, spatial, or media specialists to analyze your data and convey your research message:

- Learn how to use cutting-edge tools and methods
- Experiment with high-resolution visualization technologies
- Develop graphical representations that bring impact to your analysis

University of Virginia Library

Workshops • 1:1 Consultations • Class Presentations

Contact me at wtc2h@virginia.edu to find out more.

QUESTIONS?

Bill Corey
Data Consultant
Data Management Consulting Group
University of Virginia Library
wtc2h@virginia.edu

Andrea Horne Denton
Health Sciences Data Consultant
Claude Moore Health Sciences Library
ash6b@virginia.edu

Data Management Consulting Group
University of Virginia Library
http://dmconsult.library.virginia.edu