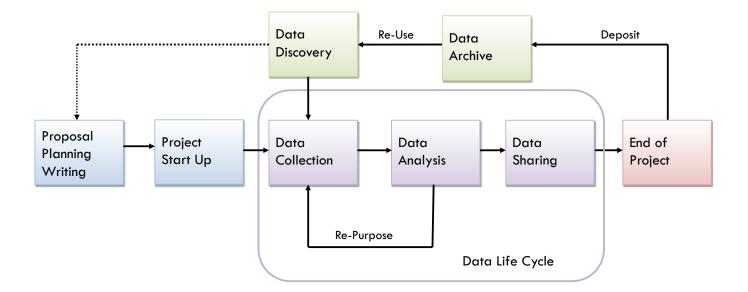
Best Practices for Collecting Data



Bill Corey Data Consultant University of Virginia Library <u>wtc2h@virginia.edu</u> Andrea Horne Denton Health Sciences Data Consultant Claude Moore Health Sciences Library <u>ash6b@virginia.edu</u>



© 2013 by the Rector and Visitors of the University of Virginia. This work is made available under the terms of the Creative Commons Attribution-ShareAlike 4.0 International license <u>http://creativecommons.org/licenses/by-sa/4.0/</u>



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

H

SUDU 30.00 80.00 80.00 80.00 80.00 80.00 80.00 Description Income: TOTAL JAN FEB MAR 1st OTR APR MAY JUN State the colspan="4">Colspan="4">Colspan="4">Colspan="4">Colspan=4" Expenses 40.00 0.00	\$0.00		JAN	FEB	MAR	1st QTR	APR	MAY	JUN						
Income: TOTAL JAN FEB MAR 1st OTR APR MAY JUN > HO COMP A COMP MAY Percent of Total:	\$0.00		\$0.00	\$0.00	\$0.00	\$0.00	\$0.00	\$0.00	\$0.00						
Income: 90.00 0.00										list-spreadsheet-contes	t-thomas-	conroy -	OpenOffic	e.org Calo	
Income: 90.00 0.00		ΤΟΤΑΙ	JAN	FEB	MAR	1st OTR	APR	MAY	JUN		(?)				
Percent of Total JAN FEB MAR 1st OTR APR MAY JUN 88,4 84 dvertising 0.00 0.00 0.00 0.00 88,4 84 <td< td=""><td>Income:</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></td<>	Income:														
Expenses: \$0.0 TOTAL JAN FEB MAR 1st OTR APR MAY JUN 88.4 I I dvertising ad Debts 0.00 0.00 0.00 0.00 88.4 84.4	a second s						0.00	0.00	0.00						_
Expenses: \$0.00 TOTAL JAN FEB MAR 1st OTR APR MAY JUN 88.4 000 dvertising ad Debts 0.00 0.00 0.00 38 964 0.00 ad Debts 0.00 0.00 0.00 38 98 98 ees 0.00 0.00 0.00 88.3 98 98 epletion 0.00 0.00 88.3 98 98 98 surance 0.00 0.00 0.00 997 287 15 15 ervices 0.00 0.00 0.00 1.5 15 15 ffice Expenses 0.00 0.00 1.2 12 12 12 epairs 0.00 0.00 0.00 0.00 0.01 0.02 0.03 0.03 eves 0.00 0.00 0.00 0.03 0.03 0.03 0.03 0.03 0.03 0.03 0.03 0.03 <t< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>E</td><td>F</td><td>G</td><td>н</td><td>1</td><td></td></t<>										E	F	G	н	1	
TOTAL JAN FEB MAR ist QTR APR MAY JUN 88.4 66.4 dvertising 0.00 0.00 0.00 38 86.4 66.5 66.5 66.5 66.5 66.5 66.5 66.5 66.5	F	to 00								Total Wt. (oz)					
IOTAL JAN FED MAK ISTOTIK AFK MAT JUN 884 884 ad Debts 0.00 0.00 0.00 0.00 38 38 ad Debts 0.00 0.00 0.00 38 38 ees 0.00 0.00 0.00 38 38 surance 0.00 0.00 62.5 38 38 surance 0.00 0.00 62.5 38 38 envices 0.00 0.00 1.5 15 15 envices 0.00 0.00 1.5 15 15 epairs 0.00 0.00 0.00 1.2 12 12 upplies 0.00 0.00 0.00 0.01 0.02 0.02 0.03 0.03 0.03 0.03 0.03 0.03 0.03 0.04 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05 <	Expenses:				1140 L	the OTD	400			88.4					
Overlap Outo	1		JAN	FEB	MAR		APR	MAY	JUN		88.4	-		-	
bes 0.00 0.00 38 apletion 0.00 0.00 88 38 anefits 0.00 0.00 62.5 surance 0.00 0.00 227 297 errors 0.00 0.00 227 297 errors 0.00 0.00 1.5 15 errors 0.00 0.00 1.5 15 errors 0.00 0.00 1.2 12 errors 0.00 0.00 1.2 12 errors 0.00 0.00 0.00 0.0 errors 0.00 0.00 0.00 0.0 errors 0.00 0.00 1.2 12 errors 0.00 0.00 0.00 0.00 errors 0.00 0.00 0.00 0.00 errors 0.00 0.00 0.00 0.00 eals 0.00 0.00 0.00 0.00 eals 0.00 0.00 0.00 0.00 eals 0.00 0.00 0.00 0.00 ther 0.00 0.00 0.00 0.00 ther 0.00 0.00 0.00		1000			-						00.4				
Pers 0.00 0.00 88 38 expletion 0.00 0.00 88 38 enefits 0.00 0.00 62.5 surance 0.00 0.00 62.5 envices 0.00 0.00 15 envices 0.00 0.00 10.00 envices 0.00 0.00 10.00 envices 0.00 0.00 12 envices 0.00 0.00 12 envices 0.00 0.00 12 upplies 0.00 0.00 0.00 vises 0.00 0.00 0.00 avel/Ent. 0.00 0.00 0.00 eals 0.00 0.00 0.00 ilities 0.00 0.00 0.00 + Neport JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 0										38					
Outo Outo Outo Outo Outo Outo enefits 0.00 0.00 0.00 62.5 900 strance 0.00 0.00 927 287 15 envices 0.00 0.00 1.5 15 15 envices 0.00 0.00 0.00 0.00 0.00 0.00 envices 0.00 0.00 1.5 15 15 generics 0.00 0.00 0.00 0.0 0.00 0.0 upplies 0.00 0.00 0.00 0.5 0.5 0.7 ave/Ent. 0.00 0.00 0.00 0.03 0.3 0.3 eals 0.00 0.00 0.00 0.03 0.3 0.3 ther 0.00 0.00 0.00 0.00 0.00 0.0 0.0 * Meport 0.00 0.00 0.00 0.00 0.0 0.0 <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>38</td> <td></td> <td></td> <td></td> <td>-</td>					-						38				-
Surance 0.00 0.00 0.00 29.7 29.7 erest 0.00 0.00 1.5 1.5 1.5 envices 0.00 0.00 0.00 0.01 0.01 fice Expenses 0.00 0.00 0.00 0.01 0.01 envices 0.00 0.00 0.00 0.01 0.01 envices 0.00 0.00 0.00 0.01 0.01 envices 0.00 0.00 0.00 0.00 0.01 envices 0.00 0.00 0.00 1.2 12 pplies 0.00 0.00 0.05 0.5 wel/Ent. 0.00 0.00 0.03 0.3 alss 0.00 0.00 0.00 0.03 ilities 0.00 0.00 0.00 0.00 eals 0.00 0.00 0.00 0.00 her 0.00 0.00 0.00 0.00 iv N/FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 0 0					-										
Surance 0.00 0.00 297 287 terest 0.00 0.00 1.5 15 envices 0.00 0.00 30.1 30.1 envices 0.00 0.00 0.00 0.00 envices 0.00 0.00 30.1 30.1 envices 0.00 0.00 1.2 12 envices 0.00 0.00 1.2 12 epairs 0.00 0.00 1.2 12 upplies 0.00 0.00 0.00 0.5 0.5 eals 0.00 0.00 0.00 0.3 0.3 ilities 0.00 0.00 0.00 0.3 0.3 eals 0.00 0.00 0.00 0.00 0.00 iher 0.00 0.00 0.00 0.					-	and the local sector				62.5					
Leferst 0.00 1.5 1.5 envices 0.00 0.00 30.1 30.1 filte Expenses 0.00 0.00 0.00 0 ent/Lease 0.00 0.00 1.2 12 epairs 0.00 0.00 1.5 0.0 upplies 0.00 0.00 1.5 0.5 ave/Ent. 0.00 0.00 0.00 0.7 eals 0.00 0.00 0.00 0.33 ilities 0.00 0.00 0.00 0.33 igges 0.00 0.00 0.00 0.00 ther 0.00 0.00 0.00 0.00 ilities 0.00 0.00 0.00 0.00 ther 0.00 0.00 0.00 0.00											297	-	1		-
Princes 0.00 301 301 fine Expenses 0.00 0.00 0 ont/Lease 0.00 1.2 12 opairs 0.00 0.00 1.2 12 opairs 0.00 0.00 1.5 opairs 0.00 0.00 0.00 opairs 0.00 0.00 1.5 opairs 0.00 0.00 0.5 0.5 opairs 0.00 0.00 0.33 0.5 was 0.00 0.00 0.33 0.3 ave/Ent. 0.00 0.00 0.33 0.3 eals 0.00 0.00 0.33 0.3 ilities 0.00 0.00 0.33 0.3 ages 0.00 0.00 0.00 0.33 her 0.00 0.00 0.00 0.00 ilities 0.00 0.00 0.00 0.00 her 0.00 0.00 0.00 0.00 ilities 0.00 0.00 0.00 0.00 ilities 0.00 0.00 0.00 0.00 ilities 0.00 0.00 0.00 0.00 ilities <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>-</td> <td></td> <td></td> <td></td>					-							-			
ent/Lease 0.00 0.00 1.2 0 epairs 0.00 0.00 1.2 12 upplies 0.00 0.00 0.00 1.5 uxes 0.00 0.00 0.5 0.5 avel/Ent. 0.00 0.00 0.7 0.7 eals 0.00 0.00 0.00 0.3 0.3 ilities 0.00 0.00 0.00 0.00 0.3 0.3 ages 0.00 0.00 0.00 0.00 0.00 0.00 0.00 h H Report JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 0 0															
epairs 0.00 0.00 1.2 upplies 0.00 0.00 1.5 uxes 0.00 0.00 0.5 avel/Ent. 0.00 0.00 0.7 eals 0.00 0.00 0.3 ilities 0.00 0.00 0.3 ages 0.00 0.00 0.00 h Report / JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 22 x Patagonia SWL/S Tshirt 65 1 65 65										0	Û				
upplies 0.00 1.5 xes 0.00 0.00 0.5 0.5 ave//Ent. 0.00 0.00 0.7 0.7 eals 0.00 0.00 0.3 0.3 ilities 0.00 0.00 0.00 0.3 0.3 ages 0.00 0.00 0.00 0.00 0.3 0.3 her 0.00 0.00 0.00 0.00 0.00 0.00 kepport JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 0 0 22 X Patagonia SWL/S T-shitt 65 1 65 65										1.2	1.2				
NXES 0.00 0.00 0.5 0.5 avel/Ent. 0.00 0.00 0.7 0.7 eals 0.00 0.00 0.3 0.3 ilities 0.00 0.00 0.3 0.3 ages 0.00 0.00 0.00 0.3 her 0.00 0.00 0.00 22.4 22 Patagonia SWL/S T-shitt 65 1 65 65															-
ave//Ent. 0.00 0.00 0.7 0.7 eals 0.00 0.00 0.3 0.3 illities 0.00 0.00 0.3 0.3 /ages 0.00 0.00 0.00 0.00 ther 0.00 0.00 0.00 0.00 * M Report JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 0 0 22 X Patagonia SWL/S Tektint 65 1 65 65										1.5					
eals 0.00 0.00 0.3 tilities 0.00 0.00 0.3 /ages 0.00 0.00 22.4 + H Report JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help 4 0 0 22 x Paragoniar SWL/ST shirt 65 1 65 65								-			0.5				
illities 0.00 0.00 'ages 0.00 0.00 ther 0.00 0.00 * H Report JAN FEB / MAY JUN JUL / AUG SEP / OCT 0 22 X Patagonia SWL/S T-shirt 65 1 65 65												1			
Ages 0.00 22.4 Prince 0.00 0 22 Prince 65 1 65 65										0.3	0.3				-
ther 0.00 22.4 H Report / JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help / 4 0 0 22 x Patagonia: SWL / ST-shirt 65 1 65 65															-
N Report / JAN / FEB / MAR / APR / MAY / JUN / JUL / AUG / SEP / OCT / NOV / DEC / Data / Help / 4 22.14 0 <td></td>															
22 x Patagonia SW L/S T-shirt 6.5 1 6.5 6.5	ther									22.4					
22 x Patagonia SWL/ST-shirt 6.5 1 6.5 6.5	H Report / JA	N / FEB / MAR /	APR (MAY ()	UN / JUL / AL	JG / SEP / OCT	r/NOV/DEC/D	ata <u>Help</u>	M		0	0				
23 x Patagonia Capilene boxer briefs 3.7 1 3.7 3.7						Patagonia SW L/S	T-shirt	6.5	1	6.5					
					23				1	3.7	3.7				
I Tarta and a sector of the se	Thursday				and the second se						9				
UNIVERSITY 24 x Low Todge Taking socias 4.5 2 9 3 VIRGINIA 25 x Patagonia synthetic socias for sandals 1.7 1 1.7 1.7	UNIVERSITY							1,7	1	1.7	1.7		11.1		

Spreadsheet Problems?

C E A B D **My Research Project** 1 Date: 5/23/2005 2 Meter Type YSI_Model_30 cond_bot 586 3 conductivity_top slack-high Tide State <30 4 5 Salinity Bottom 0.1 0.6 Top 6 Oct. 2 2005 7 Date: TIDESTATE MeterType -0.34YSI Model 30 8 9 Salinity Bottom 0.3 **Conductivity** Top 349 Salinity_Top cond_bot 39% 10 None 11 **Pause for Exercise** 12 13 Sheet1 / Sheet2 / Sheet3 I4 4 1 1 10 HH



messydata.xisx

Problems

 Dates are not stored consistently

mess	ydata.xisx					- = X
	А	В	С	D	E	F
1	My Research Pr	oject				
2	Date: 5/23/2005					
3	Meter Type	YSI_Model_30		con <mark>d_bot</mark>	58	36
4	Tide State	slack-high		conductivity_top	<30	
5	Salinity	Bottom	0.6	Тор	0	.1
6						
7	Date:	Oct. 2 2005				
8	TIDESTATE	-0.34		MeterType	YSI Mod	el 30
9	Salinity_Bottom	0.3		Conductivity Top	34	19
10	Salinity_Top	None		cond_bot	39	%
11						
12						
13						
	Sheet1 / Sheet2 / Sheet3 /	2		191		

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



Problems

 Confusion between numbers and text

🕙 mes	sydata.xisx					- = X
	А	В	С	D	E	F
1	My Research Pr	oject				
2	Date: 5/23/2005					
3	Meter Type	YSI_Model_30		cond_bot	586	
4	Tide State	s ack-high		conductivity_top	<30	
5	Salinity	Bottom	0.6	Тор	0.1	
6						
7	Date:	Oct. 2 2005				
8	TIDESTATE	-0.34		MeterType	YSI Model 3	0
9	Salinity_Bottom	0.3		Conductivity Top	349	
10	Salinity_Top	None		cond_bot	39%	
11						
12						
13						
	Sheet1 / Sheet2 / Sheet3 / Sheet3					EL.

- Different types of data are stored in the same columns
- The spreadsheet loses interpretability if it is sorted



Possible Solution

	A	В	C	D	E	F	G	
1	Date	Meter Type	TideState	Salinity_Bottom	Salinity_Top	Conductivity_Top	Conductivity_Bottom	
Z	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)

副 b	ad.xls								X		ville well_data	odbc.xls					
	E	F	G	Н	Î	J	K	L			F	F	G	Ĥ	1	1	K
1	date	meter_type	tidestate	cond_bot	cond_top	ID _	sal_bot	sal_top		4		Contractor Table Con	And the second second	COLUMN AND	research lates	erect accord	ID as
2	5/23/2005 0:00	YSI_Model_30	slack-high	586	>30	1	0.6	0.1	-	1	date	meter_type	the second second second second	weather		cond_top	IU Sa
3		й.	slack_high	268.1	273.3	2	Trace	0.2		2	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	586	237.1	1
4		0	slack high	ISSO I	1103	3	1	0.7		3	5/23/2005 0:00	YSI Model 30	slack-high	partly cloudy	268.1	273.3	2
5		1 1	clackhigh	4536	1574	4	3.2	100%		4	5/23/2005 0:00	YSI Model 30	slack-high	partly cloudy	1529	1103	3
6		0	SLACKHIG	4536	1574	- 5	3.2	1		5	5/23/2005 0:00					and the second second	and the second second
7			slack-high	<10	804	6	1000	0.5		c c	5/23/2005 0:00						
8	10/2/2005 0:00	0	falling	491	297	19	0.3	0.2		0						and the second se	3
9		0	Falling	343.6	311.	20	0.2	0.2		1_	5/23/2005 0:00	the second se		and the second sec			6
10		0	fall	2012	1316	21	1.3	0.7	D	8	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	453.3	380.7	7
11		19 C		5790	1896	22	3.7	1		9	5/23/2005 0:00	YSI Model 30	slack-high	partly cloudy	295.8	310.6	8
12		0	falling	4413	3752	23	>35ppt	2.1		In	5/23/2005 0:00	the second		and the second s		1651	9
13		n	falling	1284	635	-24	0.8	0.3		11	5/23/2005 0:00		and the second second second second	and the state of the second			10
14			Felling	350.6	353	25	0.2	0.2		10	and the second state of th	and the second sec	and the second se	and a supervision of the local distribution of the			10
15				2563	2087	26	1.6	Trace		12	5/23/2005 0:00	Constraint and of the state and the second	Contraction of the second second second	and the second	and the second s	a state of the sta	Contraction of
16		0	0	387.2	384.6	27	0.2	0.2		13	5/23/2005 0:00					2655	12
17		1. A.	falling	9010	2337	28	5.9	1.3	-	14	5/23/2005 0:00	YSI_Model_30	slack-high	partly_cloudy	338.9	330.6	13
14 4	▶ M\Sheet1 /	Sheet2 / Shee	t3 /	4	_			- P	1	15	5/23/2005 0:00	YSI Model 30	slack-high	partly cloudy	868	492	14



Possible Solution

2	A	В	C	D	E	F	G	H	1
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

c	late	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

c	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.t×t
- 🔁 Stanford-Job.pdf
- TreeRings.txt
- TreeRings_v2.txt
- Veg_Inventory
- clouddatajapan.t×t
- consistent.txt
- context.txt
- @____ dc__dec__2000_sf1_u.zip
- 🔄 filewithtwocolumns.csv
 - va-2005-stations-inv.t×t
- i word6doc.zip

File Names

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

Name

	1890_Census.txt	
10	Census.SF3.xml	
-	CloudDataJapan.csv	
	DDI-Version2-1.txt	
E	DataFile1.txt	
	DataFile2.txt	
	FlickrMetadataSynchr-v1.0.0.0.zip	
1	Library_Resources_for_Civil_and_Enviro	nmental_Engineering.pdf
	MLBS_Herb_Image_0234.jpg	
	ModelVa-Geog-orig.pl	
E	ModelVa-Geog.pl	
W	Ray_MOU_4 29 08.doc	
12	Researching_Utopian_Communities_What	at_Would_the_Librarian_Do.pd
	Safari_C30_20091010.txt	
1	Stanford-Job.pdf	
E	TreeRings.txt	
E	TreeRings_v2.txt	
	Veg_Inventory	
	clouddatajapan.t×t	
E	consistent.t×t	
E	context.txt	
	dc_dec_2000_sf1_u.zip	
	filewithtwocolumns.csv	
H	va-2005-stations-inv.txt	
18	word6doc.zip	



File Names

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

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



Organize Files Logically

Biodiversity Lake Experiments Field Work Grassland

...

...

 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
С	10.8823893	12.8
M	21.7647785	14.2
N	61.6668725	12.9
F	0.97261354	12.7
м	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)
- SSAS SAS

MATLAB

- 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) <u>example</u>
 - 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	Туре	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 <u>VisitID</u> 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	Туре	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.co m/ecomm/servlet/fsprod uctdetail_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
	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

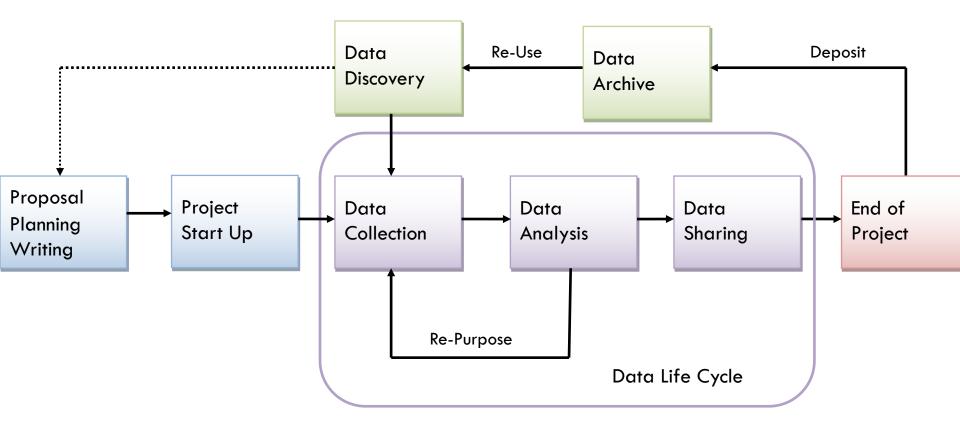


Why Manage Data?

- 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)

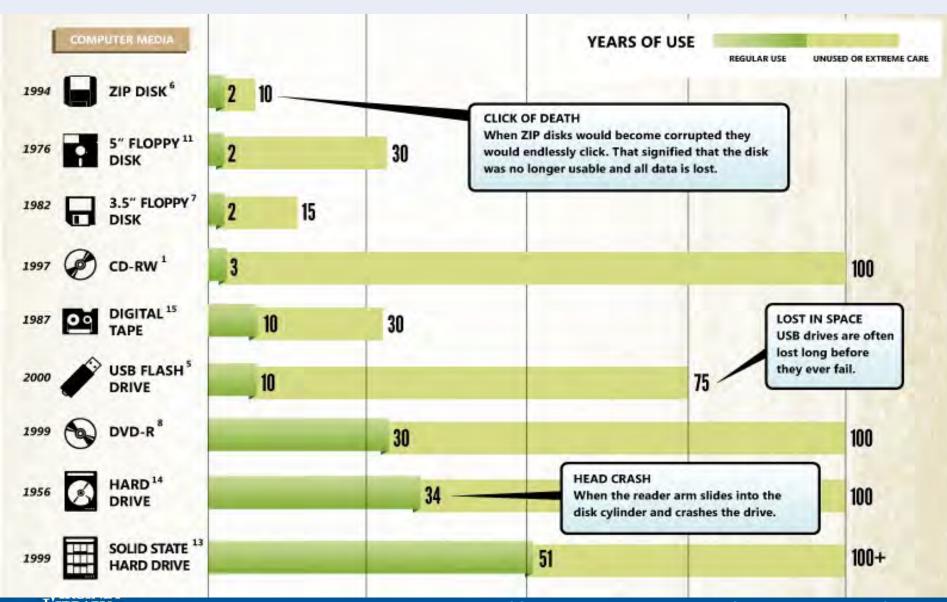


Storage & Backup





Sustainable Storage



LIBRARY

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.dataarchive.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: <u>http://eepurl.com/CJwYT</u>



More Research Data Services in the Library

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



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







옴듴움 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



QUESTIONS?

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

Andrea Horne Denton Health Sciences Data Consultant Claude Moore Health Sciences Library <u>ash6b@virginia.edu</u>

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

