

Data management

Owen Jones
jones@biology.sdu.dk



Overview

- Why you should care.
- Collecting and organising.
- Storing.
- Using.

Why you should care

- Data are the raw material of scientific studies.
- Typically 80% of effort on analysis is spent cleaning data (getting it ready to analyse).
- We should minimise this work!
- Data are valuable and should be preserved.

Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices

Species Code	B L A B I	County/Region Code	G B N K	2002	RTD Ref.
Observer Code	W A G	Locality (Post-Name)	THETFORD	Altitude	For Ringers Use Only Female Parent Age Ring No.
Day	Min.	Hour	Number of <small>(if two-letter code use column 2 see Coding Card)</small>	Status Codes	Grid Reference
15	4	9	5	FN	Male Parent Age Ring No.
22	4	13	1	HA BL FN	Young Ring Numbers
30	4	14	3	FS EY PN	
4	5	7		NN	
Comments					

HABITAT
 Refer to Nest Record Scheme Coding Card for Habitat codes.
 Choose one letter for the main habitat type (H1/H2) and then one
 number from column A. More than one number may be chosen
 from columns B and C.

FIRST HABITAT

H1 <small>One letter</small>	Column A <small>One number</small>	Column B <small>One number per box Start in left-hand box</small>	Column C <small>One number per box Start in left-hand box</small>
E	5	1 6	3

SECOND HABITAT

H2 <small>One letter</small>	Column A <small>One number</small>	Column B <small>One number per box Start in left-hand box</small>	Column C <small>One number per box Start in left-hand box</small>

NEST SITE

In	On	Under
Tree	Bridge	6.0
Bush	Earth	
Dwarf Shrub	Sand	
Creeper	Shingle	
Reeds	Stones/Rock	
Herbs	Vertical Ground	
Grass	Sloping Ground	
Dead Veg.	Flat/Gentle Slope	
Floating Veg.	Other Human Artifact	
Hedgerow	Other	
Ditch		
Wall	Nest	
Building	Centre	
	Margin	
	Field	
Give details of plant species and any extra comments on Nest Site		
FIRST HABITAT APPLE ORCHARDS		
SECOND HABITAT		
NEST SITE ON LEDGE IN BARN		
OTHER BIRD/ANIMAL NEST USED		

Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices

 NEST RECORD SCHEME CODING SYSTEM <i>Coding Card revised January 2003</i> British Trust for Ornithology, The Nunnery, Thetford, Norfolk, IP24 2PU Tel: 01842 750050, Fax: 01842 750030 Charity No. 216652	STATUS CODES <hr/> NEST BUILDING STAGE <table> <tr> <td>N0</td> <td>= Nest site empty</td> <td>N3</td> <td>= 3/4 built</td> </tr> <tr> <td>N1</td> <td>= quarter built</td> <td>N4</td> <td>= Complete, unlined</td> </tr> <tr> <td>N2</td> <td>= half built</td> <td>NL</td> <td>= Lined</td> </tr> </table> <hr/> EGGS <table> <tr> <td>CO</td> <td>= Cold</td> <td>WA</td> <td>= Warm</td> </tr> <tr> <td>UN</td> <td>= Uncovered</td> <td>CV</td> <td>= Covered</td> </tr> <tr> <td>FR</td> <td>= Fresh</td> <td>DE</td> <td>= Growing embryo present</td> </tr> <tr> <td>HA</td> <td>= Hatching</td> <td>PE</td> <td>= Pipping/calling from egg</td> </tr> </table> <hr/> YOUNG <table> <tr> <td>NA</td> <td>= Naked</td> </tr> <tr> <td>TO</td> <td>= Egg tooth present</td> </tr> <tr> <td>DO</td> <td>= Downy</td> </tr> <tr> <td>BL</td> <td>= Blind</td> </tr> <tr> <td>EY</td> <td>= Eyes just open</td> </tr> <tr> <td>IP</td> <td>= Primary feathers in pin</td> </tr> <tr> <td>FS</td> <td>= Primary feathers short; less than 1/3 emerged from sheath</td> </tr> <tr> <td>FM</td> <td>= Primary feathers medium ; 1/3 to 2/3 emerged from sheath</td> </tr> <tr> <td>FL</td> <td>= Primary feathers large; more than 2/3 emerged from sheath</td> </tr> <tr> <td>RF</td> <td>= Ready to fledge</td> </tr> <tr> <td>LB</td> <td>= Young left nest naturally before fledging; still nearby</td> </tr> <tr> <td>YR</td> <td>= Young ringed</td> </tr> <tr> <td>AY</td> <td>= Audible young in nest</td> </tr> </table> <hr/> HABITAT CODES <p>Please fill in at least Column A, and then B and C if possible. ONLY ONE CODE should be chosen from Column A, but more than one can be selected from Columns B and C.</p> <table border="1"> <thead> <tr> <th></th> <th style="text-align: center;">COLUMN A</th> <th style="text-align: center;">COLUMN B</th> <th style="text-align: center;">COLUMN C</th> </tr> </thead> <tbody> <tr> <td>A WOODLAND</td> <td> 1 Broadleaved (more than 5m tall) 2 Coniferous 3 Mixed broadleaved & coniferous (at least 10% of each) 4 Broadleaved water logged 5 Coniferous water logged 6 Mixed broadleaved and coniferous water logged </td> <td> 1 Mixed-aged or semi-natural 2 Coppice with standards 3 Coppice no standards 4 Mature plantation (taller than 10m, with closed canopy) 5 Young plantation (5-10m, open canopy) 6 Parkland (scattered trees and grassy areas) 7 High-medium disturbance from people 8 Low disturbance </td> <td> 1 Dense shrub layer 2 Moderate shrub layer 3 Sparse shrub layer 4 Dense field layer 5 Moderate field layer 6 Sparse field layer 7 Grazed (moderate to heavy) 8 Lightly grazed 9 Dead wood present 10 Dead wood absent </td> </tr> </tbody> </table> <hr/> ADULT ACTIVITY <table border="1"> <thead> <tr> <th colspan="2" style="text-align: center;">Combine (e.g. AN, PD etc)</th> </tr> <tr> <th style="text-align: center;">1st letter</th> <th style="text-align: center;">2nd letter</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;"> A = Adult M = Male F = Female P = Pair </td> <td style="text-align: center;"> D = Dead F = Feeding young at nest I = Identified by colour mark, at nest N = On/nest T = Trapped at/near nest V = In vicinity of nest B = Building nest or carrying nest material </td> </tr> </tbody> </table>	N0	= Nest site empty	N3	= 3/4 built	N1	= quarter built	N4	= Complete, unlined	N2	= half built	NL	= Lined	CO	= Cold	WA	= Warm	UN	= Uncovered	CV	= Covered	FR	= Fresh	DE	= Growing embryo present	HA	= Hatching	PE	= Pipping/calling from egg	NA	= Naked	TO	= Egg tooth present	DO	= Downy	BL	= Blind	EY	= Eyes just open	IP	= Primary feathers in pin	FS	= Primary feathers short; less than 1/3 emerged from sheath	FM	= Primary feathers medium ; 1/3 to 2/3 emerged from sheath	FL	= Primary feathers large; more than 2/3 emerged from sheath	RF	= Ready to fledge	LB	= Young left nest naturally before fledging; still nearby	YR	= Young ringed	AY	= Audible young in nest		COLUMN A	COLUMN B	COLUMN C	A WOODLAND	1 Broadleaved (more than 5m tall) 2 Coniferous 3 Mixed broadleaved & coniferous (at least 10% of each) 4 Broadleaved water logged 5 Coniferous water logged 6 Mixed broadleaved and coniferous water logged	1 Mixed-aged or semi-natural 2 Coppice with standards 3 Coppice no standards 4 Mature plantation (taller than 10m, with closed canopy) 5 Young plantation (5-10m, open canopy) 6 Parkland (scattered trees and grassy areas) 7 High-medium disturbance from people 8 Low disturbance	1 Dense shrub layer 2 Moderate shrub layer 3 Sparse shrub layer 4 Dense field layer 5 Moderate field layer 6 Sparse field layer 7 Grazed (moderate to heavy) 8 Lightly grazed 9 Dead wood present 10 Dead wood absent	Combine (e.g. AN, PD etc)		1st letter	2nd letter	A = Adult M = Male F = Female P = Pair	D = Dead F = Feeding young at nest I = Identified by colour mark, at nest N = On/nest T = Trapped at/near nest V = In vicinity of nest B = Building nest or carrying nest material
N0	= Nest site empty	N3	= 3/4 built																																																																		
N1	= quarter built	N4	= Complete, unlined																																																																		
N2	= half built	NL	= Lined																																																																		
CO	= Cold	WA	= Warm																																																																		
UN	= Uncovered	CV	= Covered																																																																		
FR	= Fresh	DE	= Growing embryo present																																																																		
HA	= Hatching	PE	= Pipping/calling from egg																																																																		
NA	= Naked																																																																				
TO	= Egg tooth present																																																																				
DO	= Downy																																																																				
BL	= Blind																																																																				
EY	= Eyes just open																																																																				
IP	= Primary feathers in pin																																																																				
FS	= Primary feathers short; less than 1/3 emerged from sheath																																																																				
FM	= Primary feathers medium ; 1/3 to 2/3 emerged from sheath																																																																				
FL	= Primary feathers large; more than 2/3 emerged from sheath																																																																				
RF	= Ready to fledge																																																																				
LB	= Young left nest naturally before fledging; still nearby																																																																				
YR	= Young ringed																																																																				
AY	= Audible young in nest																																																																				
	COLUMN A	COLUMN B	COLUMN C																																																																		
A WOODLAND	1 Broadleaved (more than 5m tall) 2 Coniferous 3 Mixed broadleaved & coniferous (at least 10% of each) 4 Broadleaved water logged 5 Coniferous water logged 6 Mixed broadleaved and coniferous water logged	1 Mixed-aged or semi-natural 2 Coppice with standards 3 Coppice no standards 4 Mature plantation (taller than 10m, with closed canopy) 5 Young plantation (5-10m, open canopy) 6 Parkland (scattered trees and grassy areas) 7 High-medium disturbance from people 8 Low disturbance	1 Dense shrub layer 2 Moderate shrub layer 3 Sparse shrub layer 4 Dense field layer 5 Moderate field layer 6 Sparse field layer 7 Grazed (moderate to heavy) 8 Lightly grazed 9 Dead wood present 10 Dead wood absent																																																																		
Combine (e.g. AN, PD etc)																																																																					
1st letter	2nd letter																																																																				
A = Adult M = Male F = Female P = Pair	D = Dead F = Feeding young at nest I = Identified by colour mark, at nest N = On/nest T = Trapped at/near nest V = In vicinity of nest B = Building nest or carrying nest material																																																																				

Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices

The screenshot shows a Google Form titled "SDU Nest Records 2013 v.2.1". The form is set against a background pattern of blue birds and flowers. At the top, there is a header with the title and a brief description: "A recording form for the SDU nest record scheme. Please complete an individual form for each nest box you have monitored." Below the header, there are several required fields:

- Recorder ID ***: A text input field with the placeholder "Use your initials (e.g. ORJ, JL etc.)".
- Box number ***: A text input field with the placeholder "Which nest box (singular) are you recording? The boxes are numbered 1 to 100 (plus "Justin")".
- Month ***: A dropdown menu with options for May, June, July, and August.
- Day ***: A text input field with the placeholder "On what day of the month did you check the box (02 = 2nd, 23 = 23rd etc.)".
- Species ***: A dropdown menu with options: BT - Blue Tit, GT - Great Tit, UK - Unknown, Other - Another bird species, and NA - Not in Use.
- Nest building status ***: A dropdown menu with options.
- Is this the second brood in this nest box?**: A question with a "Yes" radio button option.

At the bottom right of the form, there is a "Continue" button.

Collecting data

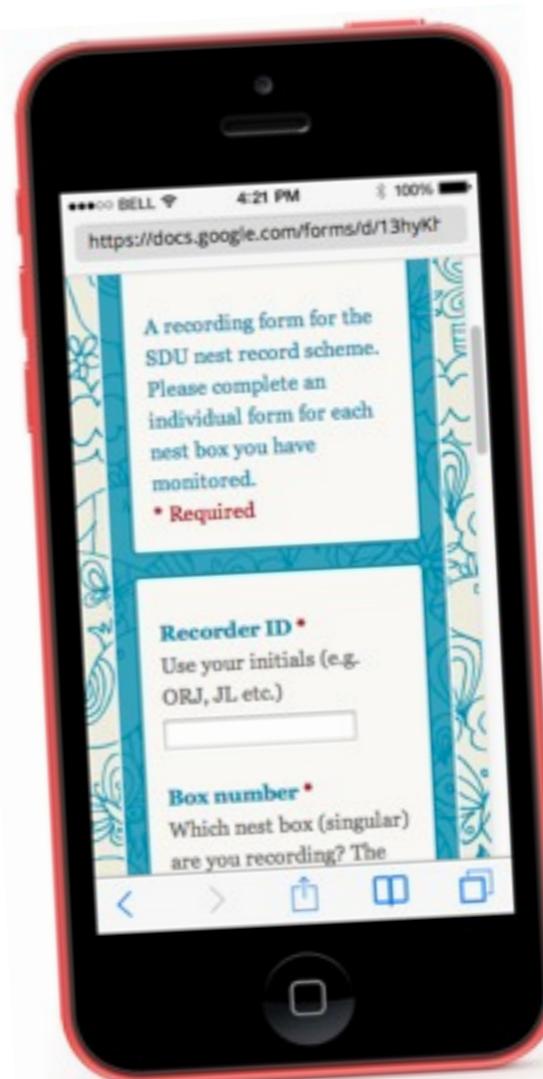
- Use printed forms/recorders
- Standardised inputs (checklist)
- Google forms
- Mobile devices

The image shows a dual-screen setup for data collection. The top screen displays the 'SDU Nest Records 2013 v.2.1' Google Form, which includes a decorative header featuring birds and flowers, and a main section with fields for 'Recorder ID', 'Timestamp', 'Box number', 'Month', 'Day', 'Species', 'Nest building status', 'Number of live eggs', 'Number of dead eggs', 'Egg status', 'Egg failure', and 'Are there any half chick'. The bottom screen shows the corresponding 'SDU Nest Record 2013 v.2.1 (Responses)' Google Sheets spreadsheet, which lists numerous rows of data collected from the form. The sheets interface includes a toolbar at the top and a table with columns labeled A through L.

Timestamp	Recorder ID	Box number	Month	Day	Species	Nest building status	Number of live eggs	Number of dead eggs	Egg status	Egg failure	Are there any half chick
5/3/2013 8:40:24	JL	1	April		9. Unk - Unknown	NL - Lined			NA - Not applicable (no eggs/cannot see eggs)		No
4/18/2013 1:50:27	TBJ and RSP	39	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:51:19	TBJ and RSP	39	April		17. Unk - Unknown	ND - 3rd built					
4/18/2013 1:52:04	TBJ and RSP	40	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:52:22	TBJ and RSP	40	April		17. Unk - Unknown	NA - quarter built					
4/18/2013 1:52:45	TBJ and RSP	41	April		9. Unk - Unknown	NA - quarter built					
4/18/2013 1:52:51	TBJ and RSP	41	April		17. Unk - Unknown	NA - Complete (lined)					
4/18/2013 1:53:27	TBJ and RSP	42	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:53:33	TBJ and RSP	42	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:54:04	TBJ and RSP	43	April		17. Unk - Unknown	ND - 3rd built					
4/18/2013 1:54:25	TBJ and RSP	43	April		9. Unk - Unknown	NA - quarter built					
4/18/2013 1:54:44	TBJ and RSP	44	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:55:04	TBJ and RSP	45	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:55:45	TBJ and RSP	46	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:55:48	TBJ and RSP	46	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:56:22	TBJ and RSP	45	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:56:41	TBJ and RSP	47	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:56:51	TBJ and RSP	46	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:57:10	TBJ and RSP	48	April		17. Unk - Unknown	ND - 3rd built					
4/18/2013 1:57:17	TBJ&RSP	47	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:57:46	TBJ&RSP	48	April		9. Unk - Unknown	NA - quarter built					
4/18/2013 1:57:58	TBJ and RSP	49	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:58:13	TBJ&RSP	49	April		9. Unk - Unknown	NA - quarter built					
4/18/2013 1:58:29	TBJ and RSP	50	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:58:41	TBJ&RSP	50	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:58:45	TBJ and RSP	51	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:59:08	TBJ&RSP	51	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:59:09	TBJ and RSP	52	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:59:32	TBJ and RSP	53	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 1:59:36	TBJ&RSP	52	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 1:59:54	TBJ and RSP	54	April		17. Unk - Unknown	ND - 3rd built					
4/18/2013 2:00:09	TBJ&RSP	53	April		9. Unk - Unknown	ND - nest site empty					
4/18/2013 2:00:23	TBJ and RSP	55	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 2:00:37	TBJ&RSP	54	April		9. Unk - Unknown	NA - quarter built					
4/18/2013 2:00:48	TBJ and RSP	56	April		17. Unk - Unknown	ND - nest site empty					
4/18/2013 2:01:03	TBJ&RSP	55	April		9. Unk - Unknown	ND - nest site empty					

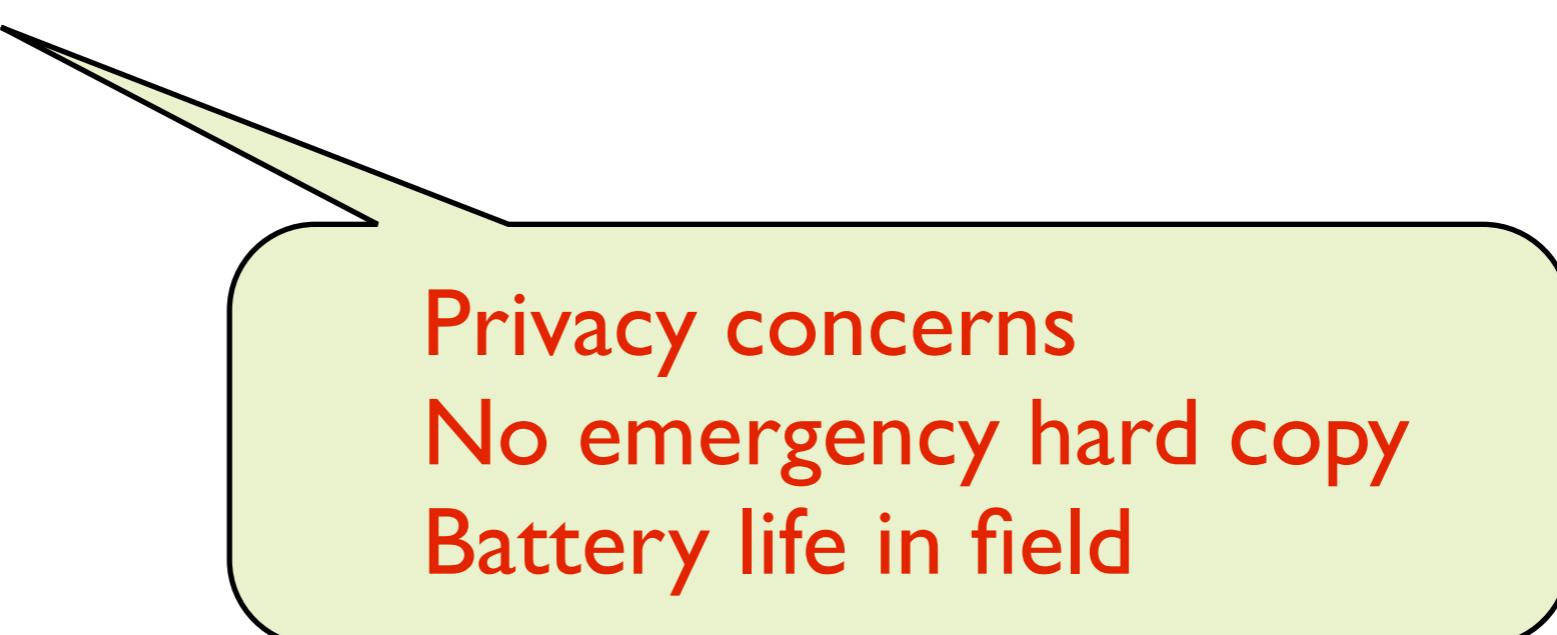
Collecting data

- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices



Collecting data

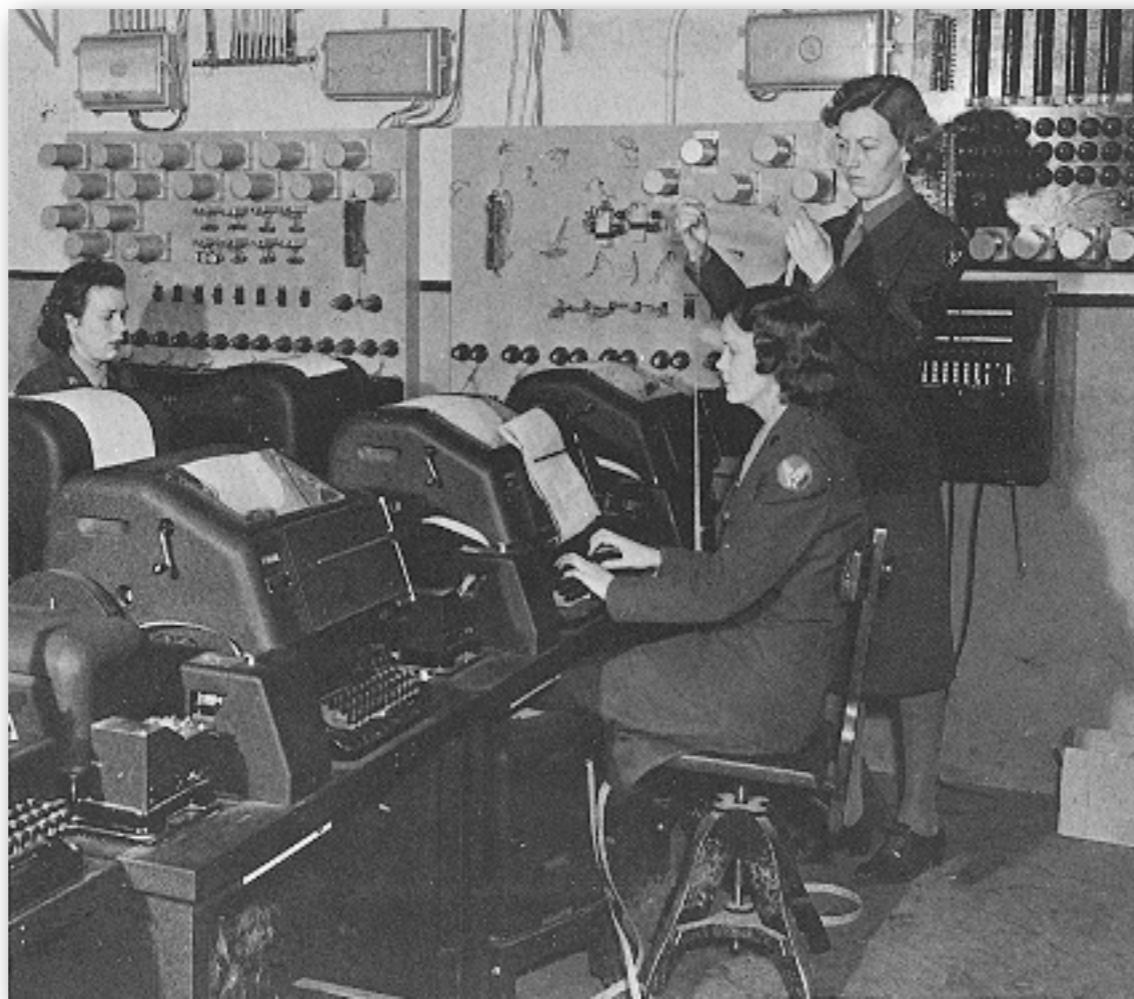
- Use printed forms/record cards
- Standardised inputs (codes)
- Google forms
- Mobile devices



Privacy concerns
No emergency hard copy
Battery life in field

Organising data

- 98%* of data entry done in Excel
- Excel is useful but can lead to bad habits



*I made this number up

Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A
1	Length
2	15
3	62
4	45
5	
6	54
7	76
8	86
9	



Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A
1	Date
2	31-Jan-76
3	27/01/82
4	27/1/82
5	14.1.70
6	05/08/1990
7	09 April 1979
8	03/04/2001
9	



Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

	A	B	C
1	ID	Length	Weight
2	1	15	3.00
3	2	62	7.00
4	3	45	4.00
5	4	did not measure	
6	5	54	9.00
7	6	76	5.00
8	7	86	7.80



Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet



A screenshot of an Excel spreadsheet illustrating several common errors:

- Row 1:** Cell A1 contains "Sample Site: Reginald Rd, Portsmouth".
- Row 2:** Cell A2 contains "Habitat: Algal".
- Row 3:** Cell A3 contains "Date collected June 16th 2010".
- Row 4:** Cell A4 contains "TrayID Tray 003".
- Row 5:** Cell A5 contains "MapRef SZ 66273 99205".
- Row 6:** Cells B6 through G6 are merged.
- Row 7:** A table starts with columns labeled "uniqueID", "Sex", "Antennal", "BodyL", and "morph".
- Row 8:** Data begins with row 8, showing entries for females and males with various morphological measurements.

	A	B	C	D	E	F	G	H
1	Sample Site: Reginald Rd, Portsmouth							
2	Habitat: Algal							
3	Date collected June 16th 2010							
4	TrayID Tray 003							
5	MapRef SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792			
6								
7		uniqueID	Sex	Antennal	BodyL	morph		
8		1	Female		6.3	11.1	Dark	
9		2	Female		7	11.8	Dark	
10		3	Female		6.1	9.3	Light	
11		4	female		5.6	9.8	Dark	
12		5	Female		5.9	11.4	Dark	
13		6	Female		5.1	9.5	Light	
14		7	Female		8.4	13.4	Light	
15		8	male		6.92	8.81	Dark	
16		9	Male		5.85	5.86	Light	
17		10	Male		5.21	6.89	Light	
18		11	Female		8.2	12	Light	
19		12	Male	NA		9.29	Dark	
20		13	male		4.88	7.38	Dark	
21		14	Female		7	9.6	Light	

Deadly sins

- Using blank columns
- Inconsistent column headers
- Merging cells
- Including multiple sheets
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth									
2	Habitat:	Algal									
3	Date collected	June 16th 2010									
4	TrayID	Tray 003									
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792					
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											



Deadly sins

- Using blank columns
- Inconsistent column headers
- Merging cells
- Including multiple sheets
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth									
2	Habitat:	Algal									
3	Date collected	June 16th 2010									
4	TrayID	Tray 003									
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792					
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											



Deadly sins

- Using blank columns
- Inconsistent column headers
- Merging cells
- Including multiple species
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- Using “sort” on only some columns

A	B	C	D	E	F	G	H	I	J	K	L
1	Sample Site:	Reginald Rd, Portsmouth									
2	Habitat:	Algal									
3	Date collected	June 16th 2010									
4	TrayID	Tray 003									
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792					
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											



Deadly sins of Excel

- Using blank cells for missing data
- Inconsistent date formatting
- Merging cells
- Including multiple tables per worksheet
- Inserting random notes
- Inserting Excel comments
- Using colour coding
- **Using “sort” on only some columns**



badExcel.xlsx

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Sample Site:	Reginald Rd, Portsmouth													
2	Habitat:	Algal													
3	Date collected:	June 16th 2010													
4	TrayID	Tray 003													
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792									
6															
7		uniqueID	Sex	Antennal	BodyL	morph									
8			1 Female	6.3	11.1	Dark									
9			2 Female	7	11.8	Dark									
10			3 Female	6.1	9.3	Light									
11			4 female	5.6	9.8	Dark									
12			5 Female	5.9	11.4	Dark									
13			6 Female	5.1	9.5	Light									
14			7 Female	8.4	13.4	Light									
15			8 male	6.92	8.81	Dark									
16			9 Male	5.85	5.86	Light									
17			10 Male	5.21	6.89	Light									
18			11 Female	8.2	12										
19			12 Male	NA	9.29										
20			13 male		4.88	7.38									
21			14 Female		7	9.6									
22			15 Male		5.97	9.44	dark								
23			16 Female		5.8	10.9	Light								
24			17 Male		3.99	7.98	Light								
25			18 Female		7.7	11	Light								
26			19 Male		6.07	8.64	Light								
27			20 Female		6.2	9.1	Light								
28			21 Female		8.1	11.6	Light								
29			22 Female		6.6	12	Dark								
30			23 Male		6.4	6.64	Light								
31			24 Male		6.07	6.81	Dark								
32			25 Male		5.47	7.76	Dark								
33			26 Female		NA	101	Light								
34			27 Female		7.7	10.7	Dark								
35			28 Male		5.75	8.59	Dark								
36			29 Male		4.06	5.33	Dark								
37			30 Male		6.35	9.65	Dark								
38			31 Male		3.12	4.79	Dark								
39			32 Male		6.05	10.01	Light								
40			33 Male		NA	6.31	Light								
41			34 Female		6.5	8.8	Dark								
42			35 Female		6.5	9.9	Light								

- Using “sort” on only some columns

Doodly-doo of Excel

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Sample Site:	Reginald Rd, Portsmouth													
2	Habitat:	Algal													
3	Date collected	June 16th 2010													
4	TrayID	Tray 003													
5	MapRef	SZ 66273 99205	Lat:	50.788635	Long:	-1.0611792									
6															
7		uniqueID	Sex	Antennal	BodyL	morph	Old Data - don't use!								
8			1 Female	3.12	11.1	Dark									
9			2 Female	3.85	11.8	Dark									
10			3 Female	3.99	9.3	Light									
11			4 female	4.06	9.8	Dark									
12			5 Female	4.88	11.4	Dark									
13			6 Female	5.1	9.5	Light									
14			7 Female	5.21	13.4	Light									
15			8 male	5.47	8.81	Dark									
16			9 Male	5.6	5.86	Light									
17			10 Male	5.75	6.89	Light									
18			11 Female	5.8	12	Light									
19			12 Male	5.85	9.29	Dark									
20			13 male	5.9	7.38	Dark									
21			14 Female	5.97	9.6	Light									
22			15 Male	6.05	9.44	Dark									
23			16 Female	6.07	10.9	Light									
24			17 Male	6.07	7.98	Light									
25			18 Female	6.1	11	Light									
26			19 Male	6.17	8.64	Light	Check this one!								
27			20 Female	6.2	9.1	Light									
28			21 Female	6.3	11.6	Light									
29			22 Female	6.3	12	Dark									
30			23 Male	6.35	6.64	Light									
31			24 Male	6.4	6.81	Dark									
32			25 Male	6.5	7.76	Dark									
33			26 Female	6.5	101	Light									
34			27 Female	6.5	10.7	Dark									
35			28 Male	6.6	8.59	Dark									
36			29 Male	6.92	5.33	Dark									
37			30 Male	7	9.65	Dark									
38			31 Male	7	4.79	Dark									
39			32 Male	7.2	10.01	Light									
40			33 Male	7.7	6.31	Light									
41			34 Female	7.7	8.8	Dark									
42			35 Female	8.1	9.9	Light									

- Using “sort” on only some columns

Good practice in Excel



- Use NA for missing data
- Dates: use columns for day, month, year
- Include only one table per worksheet
- Have a “comments” column if necessary
- Use a data column instead of colour coding
- Be very careful with “sort”

Data format

- Each variable should form a column.
- Each observation should form a row.

Bad

	treatmentA	treatmentB
John Smith	—	5
Jane Doe	1	4
Mary Johnson	2	3

Good

name	treatment	result
Jane Doe	a	1
Jane Doe	b	4
John Smith	a	—
John Smith	b	5
Mary Johnson	a	2
Mary Johnson	b	3

Data format

Problem: Column headers are values not variable names

Bad

religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k	\$75-100k
Agnostic	27	34	60	81	76	137	122
Atheist	12	27	37	52	35	70	73
Buddhist	27	21	30	34	33	58	62
Catholic	418	617	732	670	638	1116	949
Don't know/refused	15	14	15	11	10	35	21
Evangelical Prot	575	869	1064	982	881	1486	949
Hindu	1	9	7	9	11	34	47
Historically Black Prot	228	244	236	238	197	223	131
Jehovah's Witness	20	27	24	24	21	30	15
Jewish	19	19	25	25	30	95	69

Good

religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10-20k	34
Agnostic	\$20-30k	60
Agnostic	\$30-40k	81
Agnostic	\$40-50k	76
Agnostic	\$50-75k	137
Agnostic	\$75-100k	122
Agnostic	\$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Data format

Problem: Variables stored in both rows and columns

shown in Table 8. This dataset comes from the World Health Organisation, and records the counts of confirmed tuberculosis cases by country, year, and demographic group. The demographic groups are broken down by sex (m, f) and age (0–14, 15–25, 25–34, 35–44, 45–54, 55–64, unknown).

Bad

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 8: Original TB dataset. Corresponding to each ‘m’ column for males, there is also an ‘f’ column for females: f1524, f2534 and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—): this is due to the data collection process and the distinction is important for this dataset.

Data format

Problem: Variables stored in both rows and columns

Good!

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Data format

Problem: multiple data types in each column

Bad

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—	—	34.5
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—	—	16.8
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—	—	—

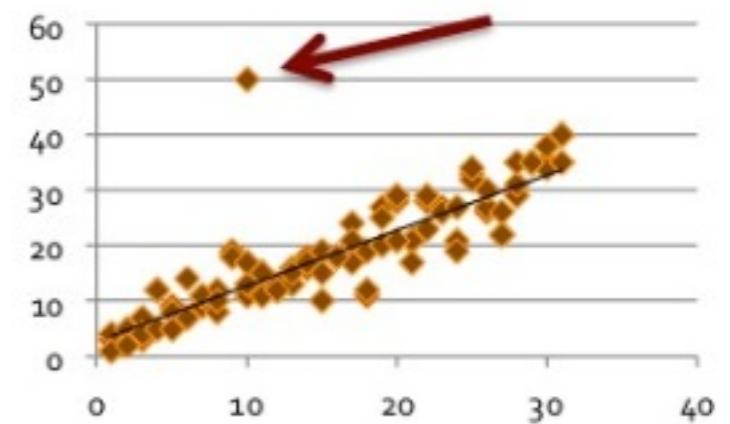
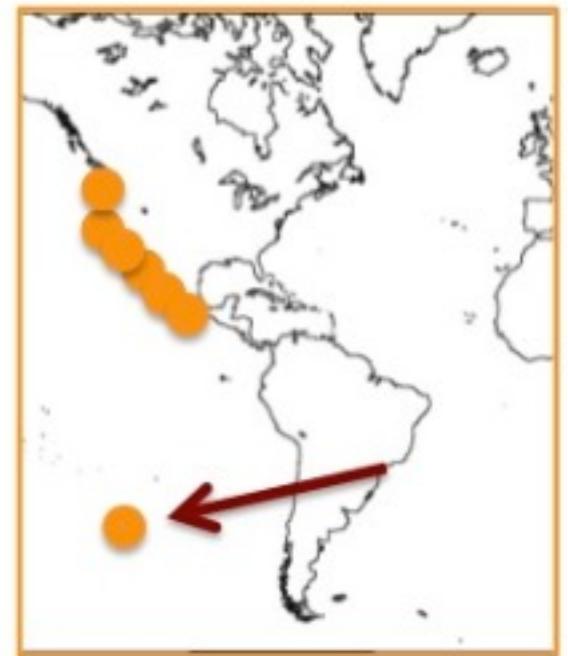
Table 10: Original weather dataset. There is a column for each possible day in the month. Columns d11 to d31 have been omitted to conserve space.

Data format

Problem: multiple data types in each column

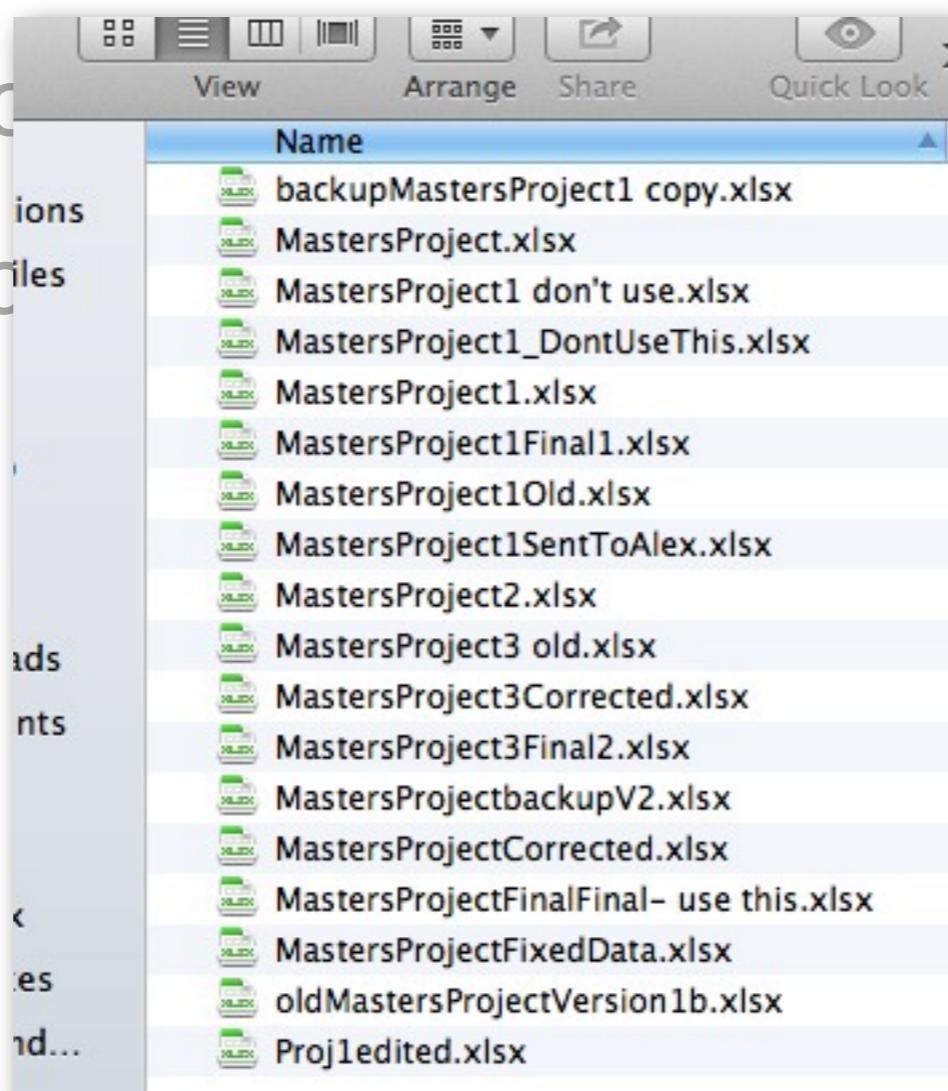
Quality control

- Check data for outliers, weird data as soon as possible.
- Aim to make work flows repeatable.
- Keep raw data raw - use scripts to process.
 - Gold standard: from raw data to analysis and plots with script(s).
 - Work is repeatable at click of a button.
 - Easy to correct.



Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable folder paths
- Backups (including versioning)



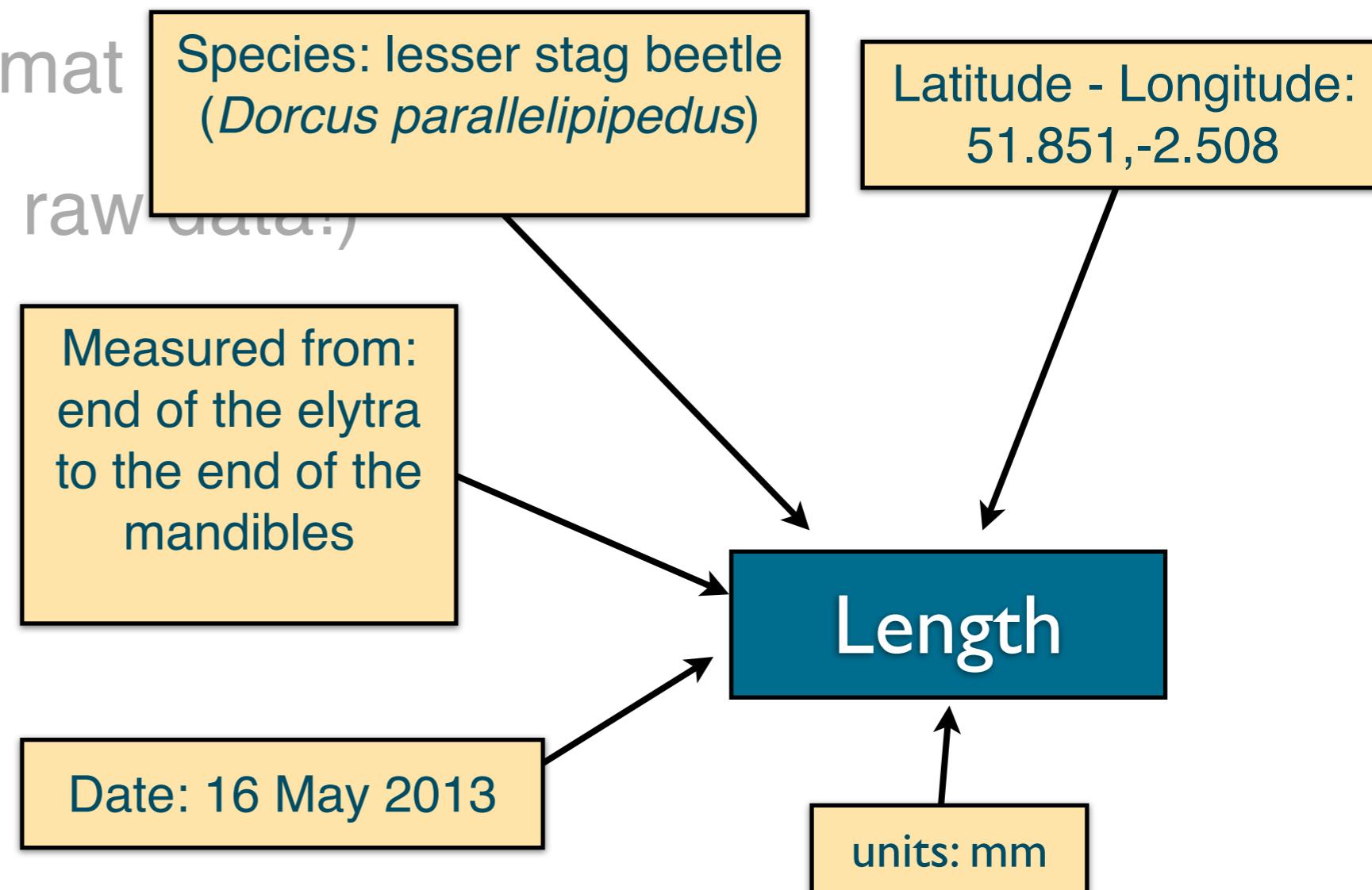
Storing data

- Use descriptive, meaningful names.
- **Include metadata**
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)

Length

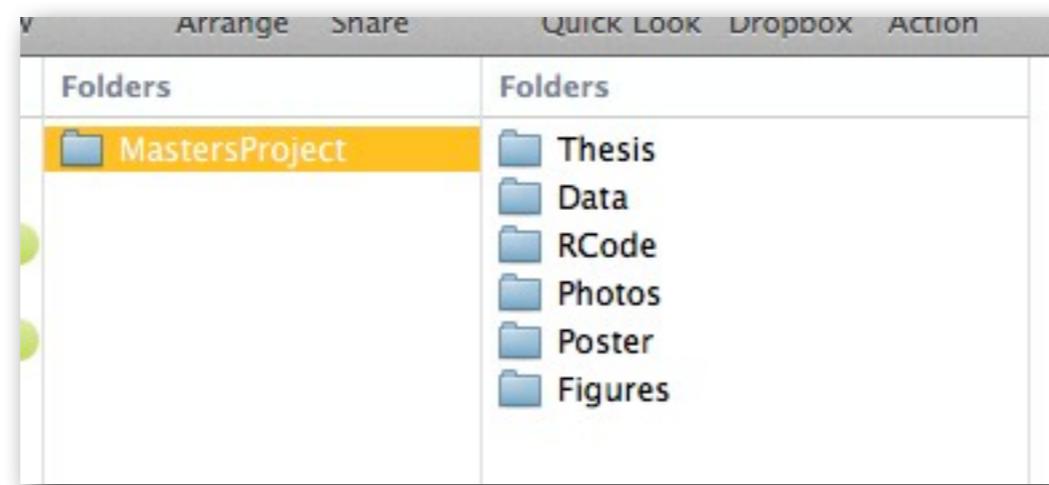
Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format
- Backups (incl. raw data.)



Storing data

- Use descriptive, meaningful names.
- Include metadata
- **Use logical file structure.**
- Use stable format (csv, txt)
- Backups (incl. raw data!)



Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)



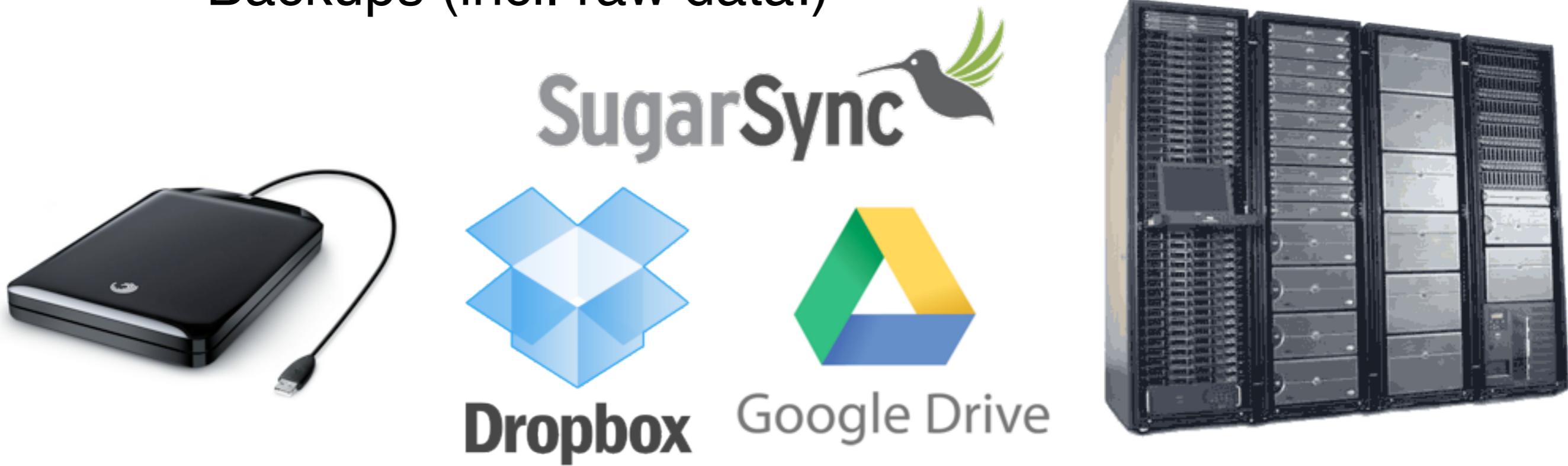
A fatal exception 0E has occurred at 0028:C00069F8 in UxD UMM<01> + 0000059F8. The current application will be terminated.

- * Press any key to terminate the current application.
- * Press CTRL+ALT+DEL to restart your computer. You will lose any unsaved information in all applications.

Press any key to continue _

Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable format (csv, txt)
- Backups (incl. raw data!)



Publish your *data*

Data increasingly recognised as a first-class research product in its own right.

Can be published.



Share this: [Share](#) 0 [Tweet](#) 1 [G+1](#) 0 [Embed*](#)

Cite this: Infection of the zebra mussel with its commensal ciliate *Conchophthirus acuminatus*. Sergey Mastitsky. figshare.
<http://dx.doi.org/10.6084/m9.figshare.95449>
Retrieved 20:21, Sep 25, 2013 (GMT)

Persistent unique ID

A screenshot of a figshare publication page. At the top, there are sharing options: Facebook Share (0), Twitter Tweet (1), Google+1 (0), and an Embed button. Below that, the citation information is shown: "Infection of the zebra mussel with its commensal ciliate *Conchophthirus acuminatus*. Sergey Mastitsky. figshare." followed by the DOI link and retrieval date. A callout bubble points to the DOI link with the text "Persistent unique ID".

Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the [Dryad Terms of Service](#). To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data. [CC ZERO](#) [OPEN DATA](#)

Title	CHONE_PC-06_LLOYDM_dataset
Downloaded	35 times
Description	This dataset was collected in the field (using a plankton pump, CTD, fluorometer and ADCP) for a MSc Thesis and 1 publication. This dataset contains the standardized abundance (count m ⁻³) of meroplanktonic larvae (Margarites spp., Crepidula spp., Astyris lunata, Diaphana minuta, Littorinimorpha, Arrhoges occidentalis, Iyanassa spp., Bittium alternatum, Nudibranchia), temperature, salinity, density, fluorescence, calculated Richardson number and current velocities (north-south, east-west, and vertical velocities) at 6 (3, 6, 9, 12, 18 and 24 m) depths, every 2 h over a 36- and 26-h sampling period, in St. George's Bay, Nova Scotia, Canada. The dataset also included lunar cycle, diel period, tidal state, date, time, sampling depth, sampling depth range, and more.

A screenshot of a Dryad dataset page. At the top, it says "Files in this package" and includes the Dryad Terms of Service. Below that is a table with three rows: "Title" (CHONE_PC-06_LLOYDM_dataset), "Downloaded" (35 times), and "Description". The "Description" row contains a detailed text block about the dataset, mentioning various species, collection methods, and sampling parameters.

Data management plans

1. Project description
2. Method description
3. Metadata format and content
4. Access and use policies
5. Long-term storage
6. Budget

See example on Blackboard

Summary

- Why you should care.
- Collecting and organising.
- Quality control/assurance
- Storing.
- Using.

To do

- Work through *QualityControl.pdf*
- Read chapter in Gotelli and Ellison.
- Read about Data Management Plans (PDFs)
- Investigate Google Forms for collecting data (see video)
- Video about metadata

Links are on Blackboard