Data management

Owen Jones
jones@biology.sdu.dk



Overview

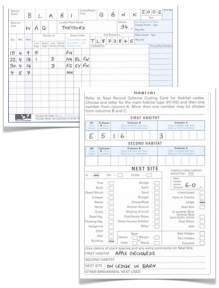
- Why you should care.
- Collecting and organising.
- Storing.
- Using.

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.

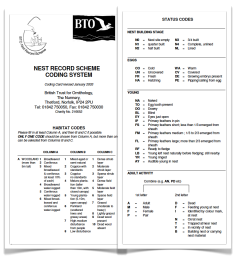
Collecting data

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



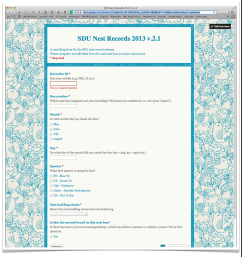
Collecting data

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



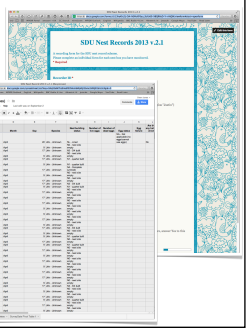
Collecting data

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



Collecting data

- Use printed forms/rec
- Standardised inputs (
- Google forms
- Mobile devices



Collecting data

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

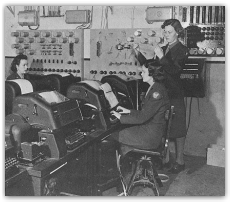
Collecting data

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

Privacy concerns
No emergency hard copy
Battery life in field

Organising data

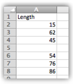
- 98%* of data entry done in Excel
- Excel is useful but can lead to bad habits



*I made this number up

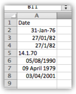
Deadly sins of Excel

- 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



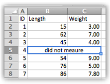
Deadly sins of Excel

- 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



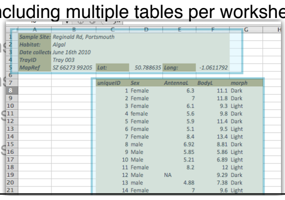
Deadly sins of Excel

- 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



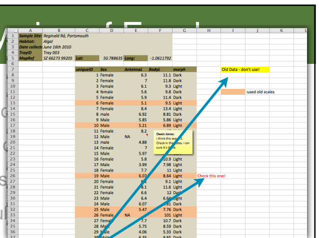
Deadly sins of Excel

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



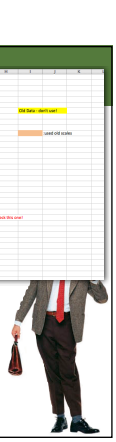

Deadly

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




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

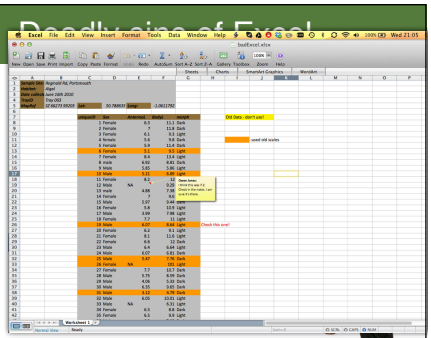


Deadly sins of Excel

- 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

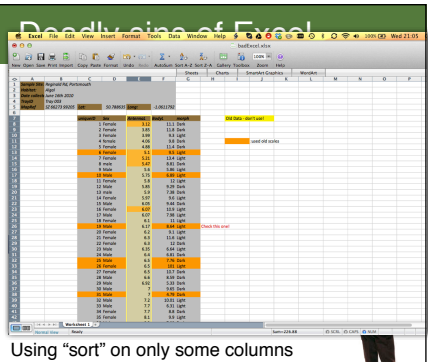
Deadly sins of Excel

- 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



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



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
DonâĂźt 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

Bad

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).

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

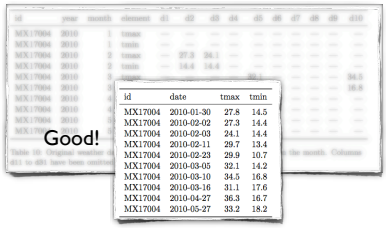
Bad

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

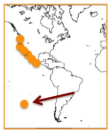


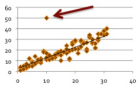
Good!

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Quality control

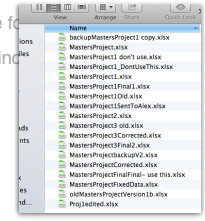
- Check data for outliers, weird data as soon as possible.
- Aim to make work flows repeatable.
- Keep raw data raw - use scripts to process.
  - Gold standard: from raw data to analysis and plots with script(s).
  - Work is repeatable at click of a button.
  - Easy to correct.





Storing data

- Use descriptive, meaningful names.
- Include metadata
- Use logical file structure.
- Use stable fo
- Backups (inc



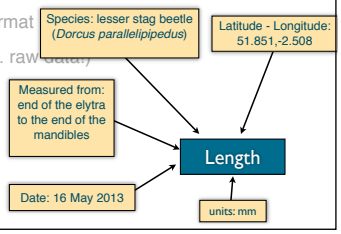
Storing data

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

Length

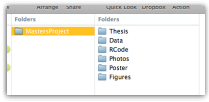
Storing data

- 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!)

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.

figshare

DRYAD

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