

# Managing Research Data Effectively

*Justin du Toit*

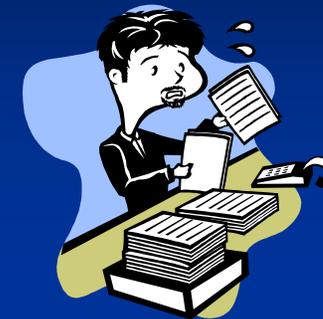
*Grootfontein Agricultural Development  
Institute*

# Entry format (base format)

Date	1 10 2008			
Site	Short grass			
	Disc Height	Number of grass plants	Number of flowering culms	Wattle seedlings
Quadrat 1	3.7 cm	6	0	12
Quadrat 2	6 cm		1	3 (1 dead)
Quadrat 3	5.3 cm	2	1	5
Quadrat 4	4.4 cm	2	2	8
Quadrat 5	6.9 cm		3	0
Date	1 10 2008			
Site	Tall grass			
	Disc Height	Number of grass plants	Number of flowering culms	Wattle seedlings
Quadrat 1	9 cm	2	2	0
Quadrat 2	11 cm	2	4	0
Quadrat 3	12.3 cm	Not recorded		
Quadrat 4	8 cm	0	0	3
Quadrat 5	8 cm	4	4	1 (dead)

- Computers view data differently from the way we do
- Computers can analyse data *incredibly* fast, but it must be in the correct format
- Data sheets are laid out in a way that makes sense to us (see left)
- Data files must be laid out in a way that makes sense to a computer

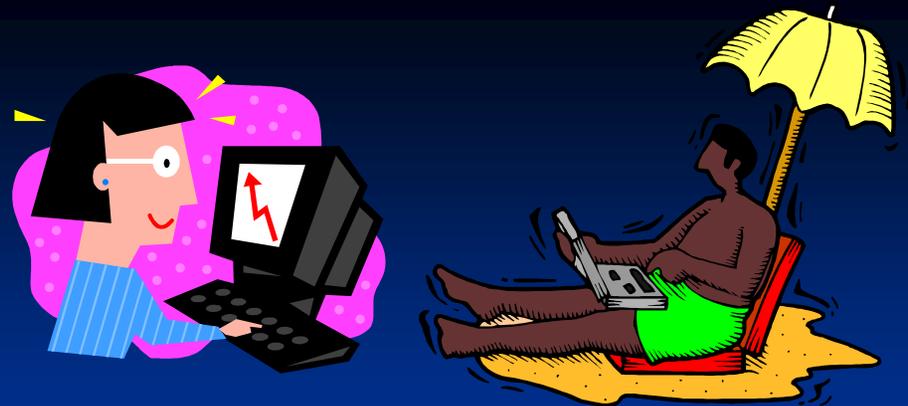
# Entry format (base format)



Date	1 10 2008			
Site	Short grass			
		Number of grass plants	Number of flowering culms	Wattle seedlings
	Disc Height			
Quadrat 1	3.7 cm	6	0	12
Quadrat 2	6 cm		1	3 (1 dead)
Quadrat 3	5.3 cm	2	1	5
Quadrat 4	4.4 cm	2	2	8
Quadrat 5	6.9 cm		3	0
Date	1 10 2008			
Site	Tall grass			
		Number of grass plants	Number of flowering culms	Wattle seedlings
	Disc Height			
Quadrat 1	9 cm	2	2	0
Quadrat 2	11 cm	2	4	0
Quadrat 3	12.3 cm	Not recorded		
Quadrat 4	8 cm	0	0	3
Quadrat 5	8 cm	4		1 (dead)

Date	Site	Quadrat number	Disc height (cm)	Number of grass plants	Number of flowering culms	Number of wattle seedlings (live)	Number of wattle seedlings (dead)
01 Oct 08	Short grass	1	3.7	6	0	12	0
01 Oct 08	Short grass	2	6	0	1	3	1
01 Oct 08	Short grass	3	5.3	2	1	5	0
01 Oct 08	Short grass	4	4.4	2	2	8	0
01 Oct 08	Short grass	5	9	0	3	0	0
01 Oct 08	Tall grass	1	9	2	2	0	0
01 Oct 08	Tall grass	2	11	2	4	0	0
01 Oct 08	Tall grass	3	12.3				
01 Oct 08	Tall grass	4	8	0	0	3	0
01 Oct 08	Tall grass	5	8	4	4	0	1

# Dual entry



- To enter a list of 100 values takes between 0.01 and 0.1% of the time that it takes to run an experiment (person working-hours, not duration of the experiment)
- However, vested within these numbers is the *entire cost* of the experiment (usually hundreds of thousands of rands)
- Therefore, it is *vital* to ensure that the data are captured perfectly
- By far the most reliable way of doing this is for two different people to enter the data (having the same person enter it twice leads to unhealthy temptation with the Copy | Paste function!)

# Spreadsheet or database?

## ■ Spreadsheets

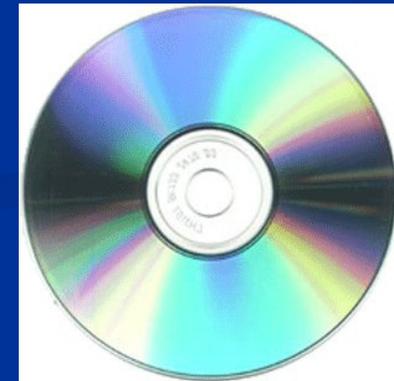
- Easy to use
- Suitable for most applications
- Lack of query capacity (hence less room for errors...)
- Good analysis and illustration capabilities
- Limited to 65536 rows x 256 columns

## ■ Databases

- Require expertise to run correctly
- VERY EASY to ask for the wrong thing, and hence end up with junk data
- Vital for large data sets, especially ongoing ones, and where >1 people enter data
- Unlimited data storage

# Backing up data

- Why backup? To have a copy of the data that is free from the hazards your original copy might face, e.g.:
  - Theft, a computer crash, losing the storage device, office burns down, etc.
- Therefore, most commonly backups face exactly the same hazards as the original!
  - Flashdrives, CDs, secondary hard-drives, laptops, etc
- The trick is to store data far, far away, and the easiest way to do that is electronically – set up a few internet-based email accounts, and Search & Send it to these regularly, and using e.g. Google Fusion
- (Similarly, photograph or scan lab- or field-data sheets)



# Data accessibility and online-storage

- Many researchers control many valuable data sets (much of it historical) that have cost lots of money to produce
- Much of this could easily be made accessible to others to use (don't be miserly! – may also initiate a collaboration with other researchers).
- More and more data lost as people jump between jobs
- Storing data remotely is useful

# Example – Google Fusion

The image displays two side-by-side browser windows showing Google Fusion Tables. Both windows are titled "No place Justin du Toit DAFF" and show a table of data. The left window displays a table with columns: Date, Transect, Longitude, Latitude, and Altitude. The right window displays a table with columns: Date, Full species, Genus, Species, and Variety. Both windows show data for the date 5/13/2010.

Date	Transect	Longitude	Latitude	Altitude
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	1	-31.257804	25.503136	1400
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385
5/13/2010	2	-31.260075	25.496845	1385

Date	Full species	Genus	Species	Variety
5/13/2010	Aristida congesta	Aristida	congesta	
5/13/2010	Eragrostis lehamanniana	Eragrostis	lehamanniana	
5/13/2010	Aristida diffusa	Aristida	diffusa	
5/13/2010	Eriocephalus spinescens	Eriocephalus	spinescens	
5/13/2010	Walafrida saxatilis	Walafrida	saxatilis	
5/13/2010	Tragus koeleroides	Tragus	koeleroides	
5/13/2010	Pentzia globosa	Pentzia	globosa	
5/13/2010	Eragrostis curvula	Eragrostis	curvula	
5/13/2010	Chrysochoma ciliata	Chrysochoma	ciliata	
5/13/2010	Eriocephalus ericoides	Eriocephalus	ericoides	
5/13/2010	Helichrysum pentzoides	Helichrysum	pentzoides	
5/13/2010	Zygophyllum incrustatum	Zygophyllum	incrustatum	
5/13/2010	Eragrostis curvula	Eragrostis	curvula	
5/13/2010	Aristida diffusa	Aristida	diffusa	
5/13/2010	Cymbopogon plurinodis	Cymbopogon	plurinodis	
5/13/2010	Nanav micronhulla	Nanav	micronhulla	

# Example – Google Fusion

The merged table will be computed dynamically from the two base tables, using the rows from the first table. Changes to the base tables will be reflected in the merged table and vice versa.

Merge tables Cancel

Current view: All - Show options

Location Latitude  Display as heat map [Configure info window](#) [Configure styles](#) [Export to KML](#) [Get KML network link](#) [Get embeddable link](#)

Map	Satellite	Hybrid	Terrain
-----	-----------	--------	---------

Date: 5/18/2010  
Transect: 6  
Longitude: -31.24893  
Latitude: 25.512645  
Altitude: 1421  
Full species: Eriocephalus ericoides  
Genus: Eriocephalus  
Species: ericoides  
Variety: Erii  
Code: Erii  
Number: 1  
Grazing: Grazed

©2010 Google - Map data ©2010 Tele Atlas - Terms of L

start 2 Microsoft Office ... 5 Microsoft Office ... 2 Internet Explorer EN 10:12 AM Thursday

# Make your life easier in Excel! Learn the basic functions!

- Basic arithmetic functions (=, -, x, /, sum, average, count)
- Simplify data – e.g. convert continuous data to binomial (=if(VALUE>x,1,0))
- Transform data – e.g. (=log(VALUE+1))
- Look up information about a common value or term (=vlookup(VALUE, TABLE, COLUMN))
- Remove unwanted spaces (=trim(TEXT))
- Conditionally manipulate data (=if(CONDITION, TRUE, FALSE))
- Split data into separate records e.g. “Eragrostis curvula” into two columns titled “Genus” and “Species”
- Truncate text e.g. (=left(“Eragrostis”,3) gives “Era” – combined with the same for “curvula” (cur) can be joined (=“Era”&“cur”) to give a code **Eracur**
- You NEVER have to re-enter values, and NEVER have to use a calculator in Excel