Science & Technology Facilities Council | UK Research and Innovation | NERC SCIENCE OF THE ENVIRONMENT

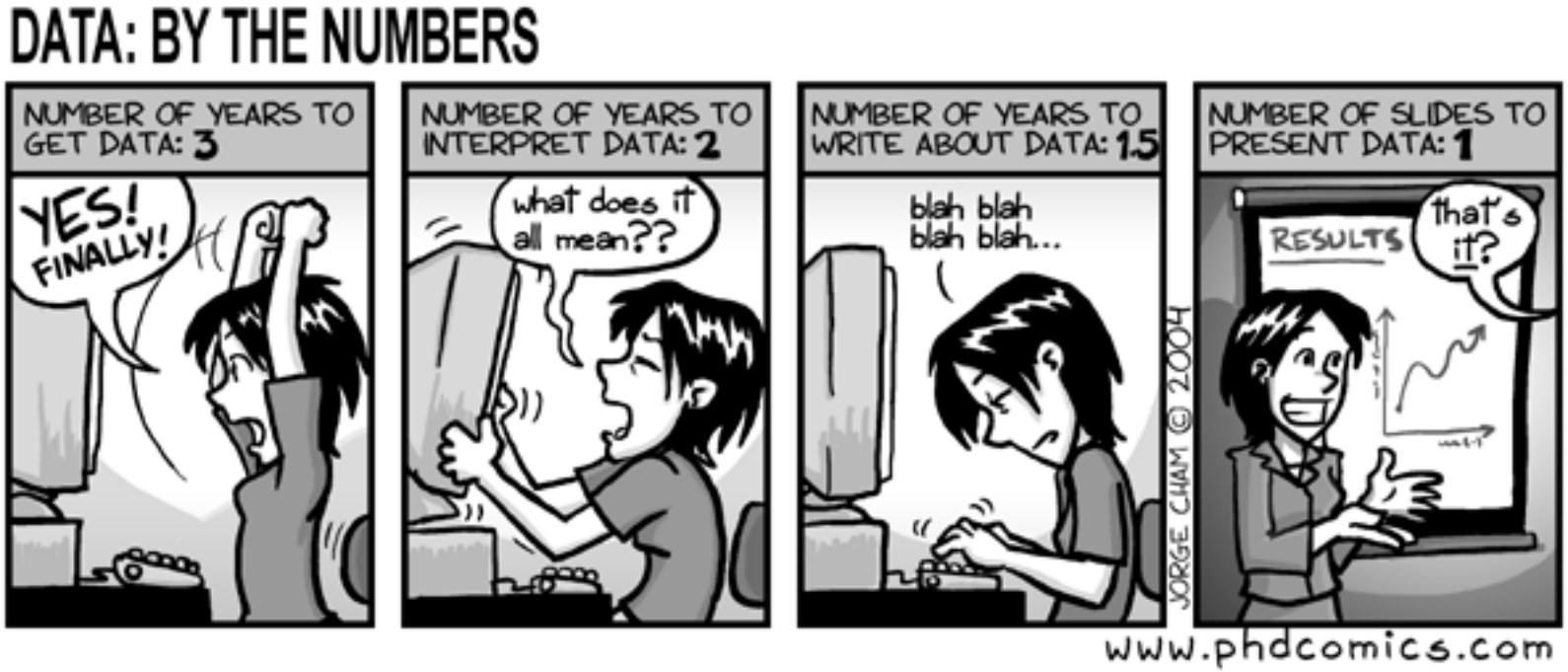
Why Good Data Curation is Essential for Doing Good Science

Alison Pamment

On behalf of the course team

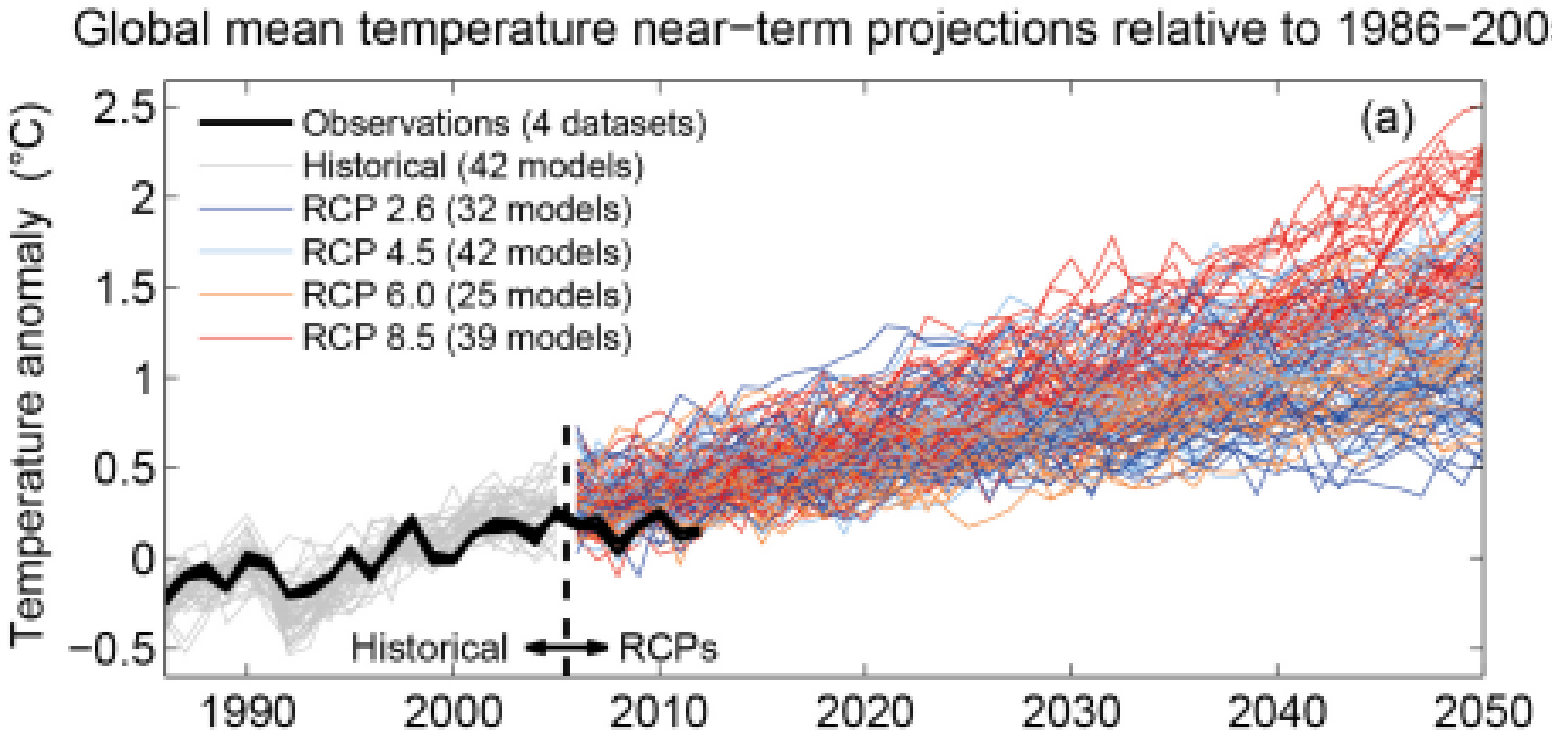
(NCAS/NCEO CEDA, NCAS CMS, NCAS Operations)

Creating a dataset is hard work!



"Piled Higher and Deeper" by Jorge Cham www.phdcomics.com

IPCC global mean surface temperature projections



IPCC Working Group I, 5th Assessment Report, Chapter 11, Figure 25.

Reasons to care about good data management (1)

- Data can be expensive, even impossible, to reproduce
- As scientists we need to be able to analyse and re-analyse our data
- We need a systematic, automated approach to handle large data volumes
- We need to share our data with collaborators
- We want to compare with data produced by other researchers

Reasons to care about good data management (2)

- We want to archive our data for long term preservation (often a funding requirement)
- We want our work to be cited in other studies to gain academic credit
- **We need robust and efficient methods of reading, writing, storing, moving, finding and citing data**

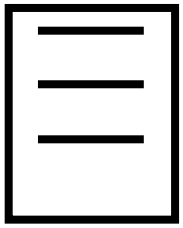
Automating data interactions

- Wherever possible we use:
  - ➡ common software tools
- which are designed to work with
  - ➡ standard file formats
- which in turn comply with
  - ➡ metadata conventions
- **It takes effort to learn these…**

  … but they make your life easier in the end

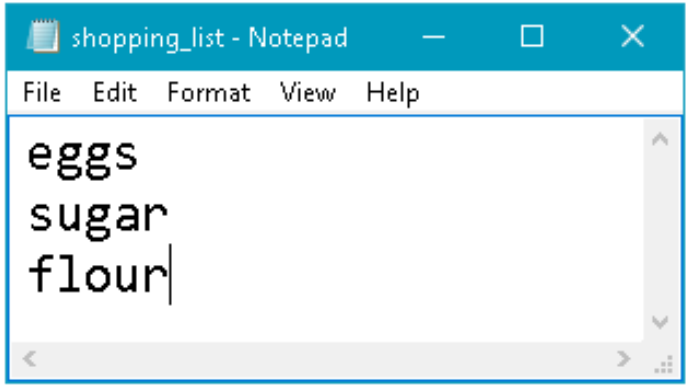
Standard file formats

Standard file formats: ascii text



shopping_list.txt

```
$ cat shopping_list.txt
eggs
sugar
flour
```



An ascii file seems simple and standard computer operating systems, e.g. Linux, Windows, Mac, etc. allow the user to easily create and display such files but it is still a **binary encoded file format**

ASCII Encoding table

Codes for:
Lower case a – z
Upper case A-Z
Digits 0 – 9
Punctuation !?()% etc.
Terminal control

Inside the file:
85 (53 in hex,
01010011 bin) is “S”.

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

Ascii extensions

- ASCII encoding has been more or less the same since 1963.
- Unicode (UTF-8) retains the original ascii codes but extends to many thousands of characters

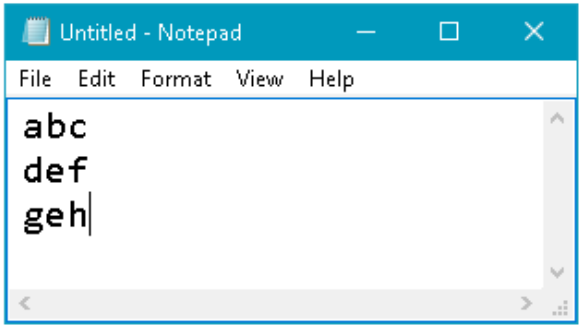
ñ	241	00F1	ñ	LATIN SMALL LETTER N WITH TILDE
ò	242	00F2	ò	LATIN SMALL LETTER O WITH GRAVE
ó	243	00F3	ó	LATIN SMALL LETTER O WITH ACUTE
Α	913	0391	Α	GREEK CAPITAL LETTER ALPHA
Β	914	0392	Β	GREEK CAPITAL LETTER BETA
Γ	915	0393	Γ	GREEK CAPITAL LETTER GAMMA
Δ	916	0394	Δ	GREEK CAPITAL LETTER DELTA
☼	9788	263C		WHITE SUN WITH RAYS
☽	9789	263D		FIRST QUARTER MOON
☾	9790	263E		LAST QUARTER MOON

Text file gotchas

- Some text characters are not represented in Unicode, e.g., ‘smart’ quotes: “ ” ‘ ’ and ‘non-printing’ characters – most often this results from copying and pasting from word processors and web browsers

➡ **Use text editor settings to save as ascii or unicode**

- Moving between operating systems, e.g. Linux and Windows



→

```
$ vi myfile.txt
abc^M
def^M
geh^M
```

➡ **Use unix2dos or dos2unix Linux commands to add/remove ^M characters at end of line**

Ascii for data storage

- Useful for small amounts of data, e.g. list of values, and for temporary storage
- Portable and many tools available for reading / writing
- It's a standard file encoding, but no standard way to structure data so difficult to develop standardized data processing tools
- No guarantee that metadata will exist or be structured
- **Not recommended for long-term data storage**

Comma separated variables

- Structured ascii file that can be used with a standard text editor so it is ‘human readable’
- Easily processed by spreadsheet applications (data arranged in rows and columns)
- Defined csv format for atmospheric measurements: **BADC-CSV (.csv)**
- This data storage format includes defined **metadata conventions** and **file structure.**

BADC-CSV (1)

	A	B	C	D
1	Conventions	G	BADC-CSV	1
2	title	G	My data file	
3	creator	G	Prof W E Ather	Reading
4	contributor	G	Sam Pepler	BADC
5	creator	G	A. Pdra	
6	long_name	1	time	days since 2007-03-14
7	long_name	2	air temperature	
8	long_name	3	met station air temperature	
9	creator	3	unknown	Met Office
10	coordinate_variable	1	x	
11	location_name	G	Rutherford Appleton Lab	
12	data			
13	1	2	3	
14	0.8	2.4	2.3	
15	1.1	3.4	3.3	
16	2.4	3.5	3.3	
17	3.7	6.7	6.4	
18	4.9	5.7	5.8	
19	end data			
20				

BADC-CSV (2)

$ cat simple-example.csv

```
Conventions,G,BADC-CSV,1
title,G,My data file,
creator,G,Prof W E Ather,Reading
contributor,G,Sam Pepler,BADC
long_name,1,time, days since 2007-03-14
long_name,2,air temperature,
long_name,3,met station air temperature,
creator,3,unknown,Met Office
coordinate_variable,1,x,
location_name,G,Rutherford Appleton Lab,
data,,,
1,2,3,
0.8,2.4,2.3,
1.1,3.4,3.3,
2.4,3.5,3.3,
3.7,6.7,6.4,
end data,,,
```

CSV gotchas

- Non-ASCII characters – cut and paste from other places
- Date / time formats can vary between applications and formats,
  e.g. BADC-CSV expects YYYY-MM-DD hh:[mm:ss]
- Coordinate values should be monotonic
- ‘Missing data’ value must be outside valid range of data
- Possible to write metadata that are no help, e.g. author’s name = Sam, variable name = sam3.

BADC-CSV for data storage

- Works well for storing 1D data
- Portable and many tools available for reading / writing (although still subject to ascii 'gotchas')
- Data and metadata standards are defined – use them!!
- Examples and documentation: http://help.ceda.ac.uk/category/4423-formats
- **Provided standards are followed, BADC-CSV is suitable for long-term data storage**

NetCDF (.nc)

- For big data need more flexible file formats such as NetCDF
- Not ascii, therefore not directly human readable
- Efficient for storing large volumes of numeric data – byte, integer, floating point data types
- Can store multi-dimensional arrays, e.g. NumPy arrays
- Portable (independent of hardware architecture)
- Many software tools available to process NetCDF files and manipulate data
- Metadata conventions well defined, most notably, NetCDF conventions and CF conventions

Other file formats

Many file formats are in use:

NASA-Ames (.na) ASCII format (not comma separated)

GRIB (.grb) GRIdded Binary: WMO operational format for models

PP (.pp) Output from Met Office Unified Model

.png, .jpg Image file formats

.mp4 Video file format

xls, docx, etc. Proprietary formats from common software tools

Steer clear of proprietary and bespoke formats for long term data storage!!

Metadata

Metadata – Data about the data

Who produced the data?

How was it done?

Why was it done?

When was it produced?

Where does the data relate to?

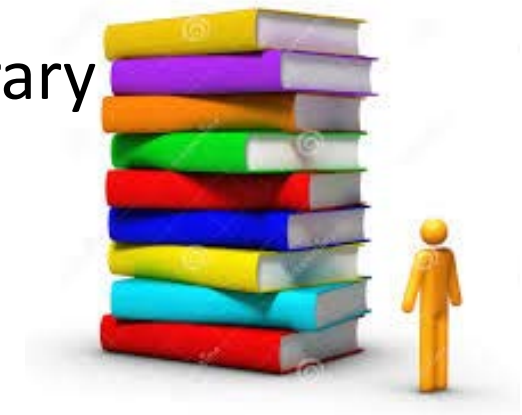
What are the data?

Metadata – Data about the data

- **Discovery metadata** – enable the data to be found, e.g. experiment name, date, geographical area
- **Browse metadata** – more detailed metadata, e.g., what variables were observed/modelled
- **Usage metadata** – highly detailed e.g. variable names, units, precise coordinates, processing algorithms
- **Citation metadata** – e.g. links to academic papers citing the data, post fact annotations
- **'Extra' metadata** – e.g. detailed metadata about the instrument used

Discovery metadata (1)

I want to find a library book on Python programming…



…I can search the library catalogue for “python”…

Discovery metadata (2)

"Monty Python at Work", Michael Palin. Publisher: Hern Books. TV Comedy.

"Learning Python", Mark Lutz. Publisher: O' Reilly. Computer Programming.

"Ball Pythons: Caring For Your New Pet (Reptile Care Guides)", Casey Watkins. Publisher: TokaySEO. Animal care.

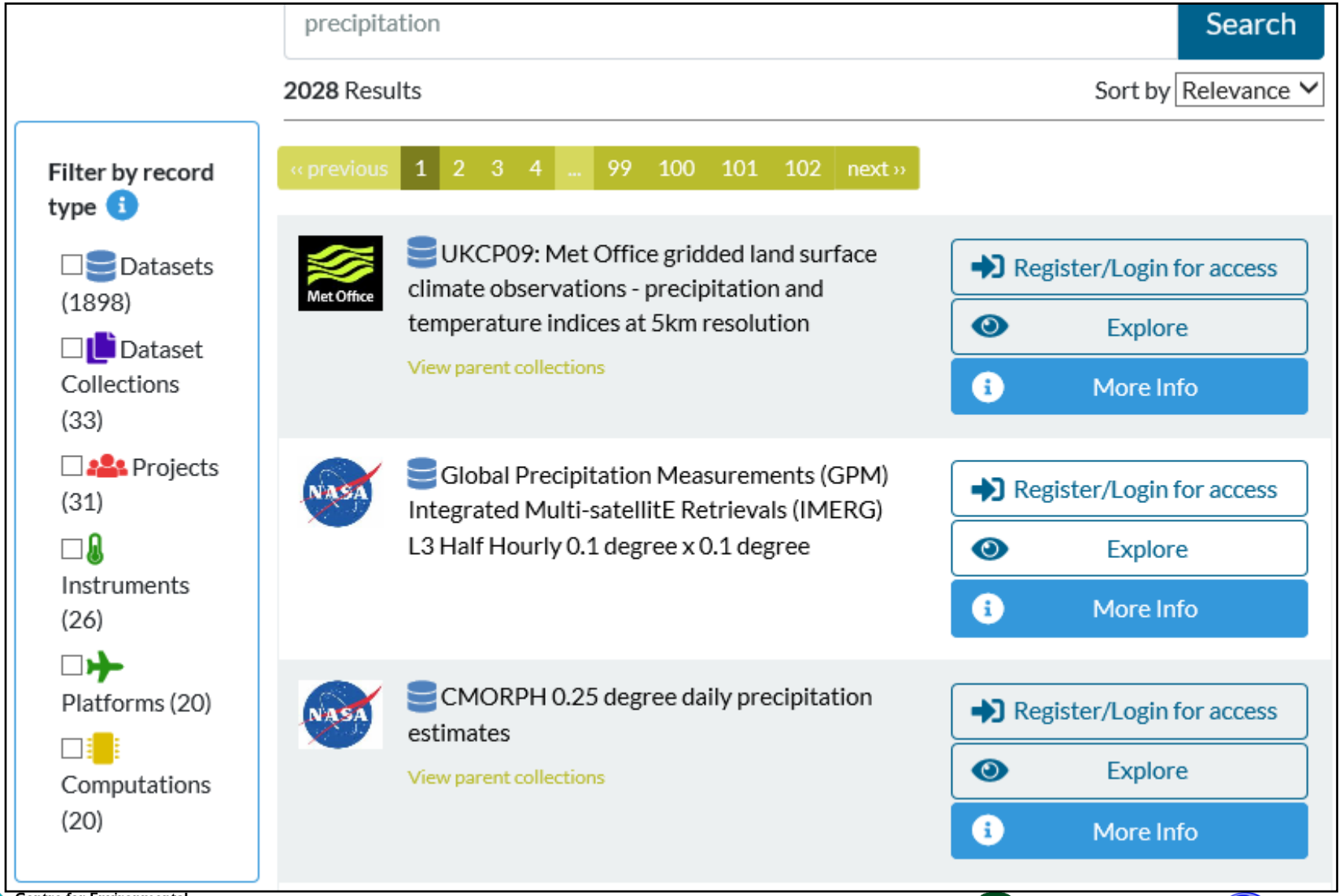
Discovery metadata (3)

… I refine my search to “Learn python” …

“Learning Python”, Mark Lutz.
Publisher: O’ Reilly.

2015, 382 pp, Computer Science, Shelf Mark 3L52, Dewey: 00532.44.3

http://catalogue.ceda.ac.uk



Usage Metadata

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What is known about this file?

sw indicates that the file contains "surface" wind data (i.e. speed and direction) from the location Frongoch

010203 represents the date in YYMMDD format

1st February 2003
(British convention)

2nd January 2003
(North American convention)

3rd February 2001
(Swedish convention)

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

What can we guess?

- Values are clearly arranged in pairs
  - 1st value of pair (e.g. 4.31) must represent speed - probably in units of m s^{-1}
  - 2nd value of pair (e.g. 155.3) must represent direction - probably in units of ⌖ from North (but meteorological or vector convention?)
- 240 lines, each with 6 columns, each with a pair of values => 1440 pairs of values
- There are 1440 minutes in a day => 1 minute sampling

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

In which order should we read the data?

Column by column and then row by row or *vice versa*?

Try both ways and plot time series of the speed and direction data

There should be no sharp discontinuities in speed or direction

Vector (i.e. towards which the wind is blowing) or meteorological direction?

Compare with synoptic pressure maps or MST radar data

The first 10 (of 240) lines from the file **sw010203**
(taken from the NERC MST Radar Facility archives)

```
4.31 155.3  3.92 136.1  5.15 140.2  4.23 137.1  4.75 150.2  4.71 137.9
4.35 146.5  4.52 138.0  4.83 153.7  5.40 145.8  4.63 141.0  4.90 137.3
4.31 143.3  4.58 157.0  4.94 141.7  4.65 143.1  4.63 143.0  4.88 149.5
5.42 148.5  4.92 140.4  4.04 146.7  3.92 151.5  5.02 135.3  5.06 151.6
4.65 152.3  4.31 168.8  3.79 145.3  5.92 152.9  5.02 145.8  4.77 161.6
4.79 144.1  4.60 147.5  5.33 150.1  4.81 141.0  6.02 146.9  4.38 149.0
4.42 142.5  4.58 133.4  4.35 150.5  4.96 149.8  5.56 143.4  5.08 148.5
5.19 141.6  4.40 142.4  4.10 152.6  5.02 134.0  4.94 142.9  5.27 144.4
5.38 141.5  5.88 144.8  6.00 140.1  4.75 158.3  5.08 148.1  5.46 163.5
4.27 150.8  4.69 138.8  5.71 144.0  5.21 138.8  5.00 132.4  5.06 144.4
```

It is often possible to "decode" ASCII files in this way, it is much more difficult for binary.

No-one will be prepared to make this effort unless they have a strong need for the data.

The data will becomes useless if the file name is changed - the date information is not recorded anywhere else.

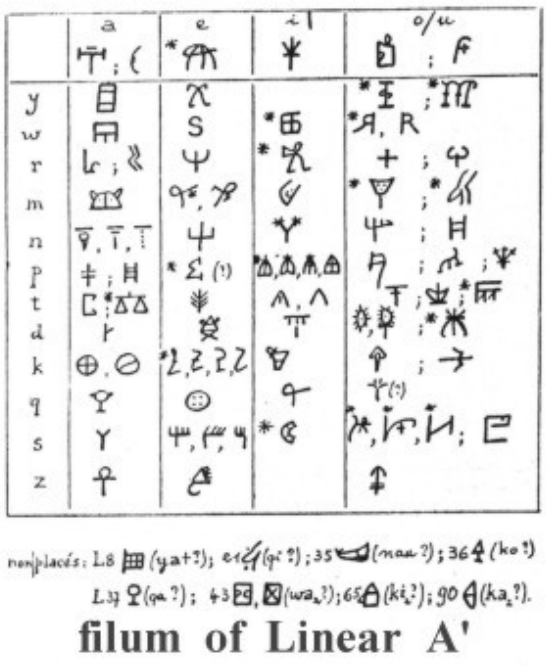
Even if the data can be read, they may be of little scientific value unless something is known about: the type of instrument used, where it was located & how it was operated.

The partial contents of file nerc-mstrf-wind-sensors_capel-dewi_20080114_wxt510.nc

```
global attributes:
          :verbose_metadata = "Free text description" ;
          :file_version_number = 1s ;
          :data_year = 2008s ;
          :data_month = 1s ;
          :data_day = 14s ;
dimensions:
          time = 1440 ;
variables:
          float longitude() ;
                    longitude:units = "degrees_east" ;
                    longitude:axis = "X"
          float latitude() ;
                    latitude:units = "degrees_north" ;
                    latitude:axis = "Y" ;
          float altitude() ;
                    altitude:units = "m" ;
                    altitude:axis = "Z" ;
          int time(time) ;
                    time:units = "seconds since 2008-01-14 00:00:00 +00:00" ;
                    time:axis = "T" ;
          float mean_wind_speed(time) ;
                    mean_wind_speed:units = "m s-1" ;
                    mean_wind_speed:coordinates = "latitude longitude altitude" ;
                    mean_wind_speed:cell_methods = "time: minimum (interval: 3 s)" ;
                    mean_wind_speed:missing_value = 99.9f ;
          short mean_wind_direction(time) ;
                    mean_wind_direction:units = "degree" ;
                    mean_wind_direction:coordinates = "latitude longitude altitude" ;
                    mean_wind_direction:cell_methods = "time: minimum (interval: 3 s)";
                    mean_wind_direction:missing_value = 999s ;
```

But why go to all this trouble?

It’s ok, I’ll just do regular backups



filum of Linear A'

Phaistos Disk, 1700BC

These documents have been preserved for thousands of years! But they’ve both been translated many times, with different meanings each time.

Data preservation is not enough: we need to actively curate to preserve Information.

Increasing Data Impact

Good data and metadata formats...

- Help to guarantee unambiguous content
- Permit metadata harvesting from the data
- Ensure future users can open data files
  - How future proof is an Excel spread sheet?
- Enable data to be cited (DOI)
- And let the scientists concentrate on doing science

How NOT to manage your data…

https://www.youtube.com/watch?v=N2zK3sAtr-4