How ***NOT*** to share your data:

Avoiding data horror stories

Rosie Higman
Office of Scholarly Communication

8th March 2017

1. Where?
2. What?
3. File formats
4. Formatting your spreadsheet*
5. Document and describe your data!

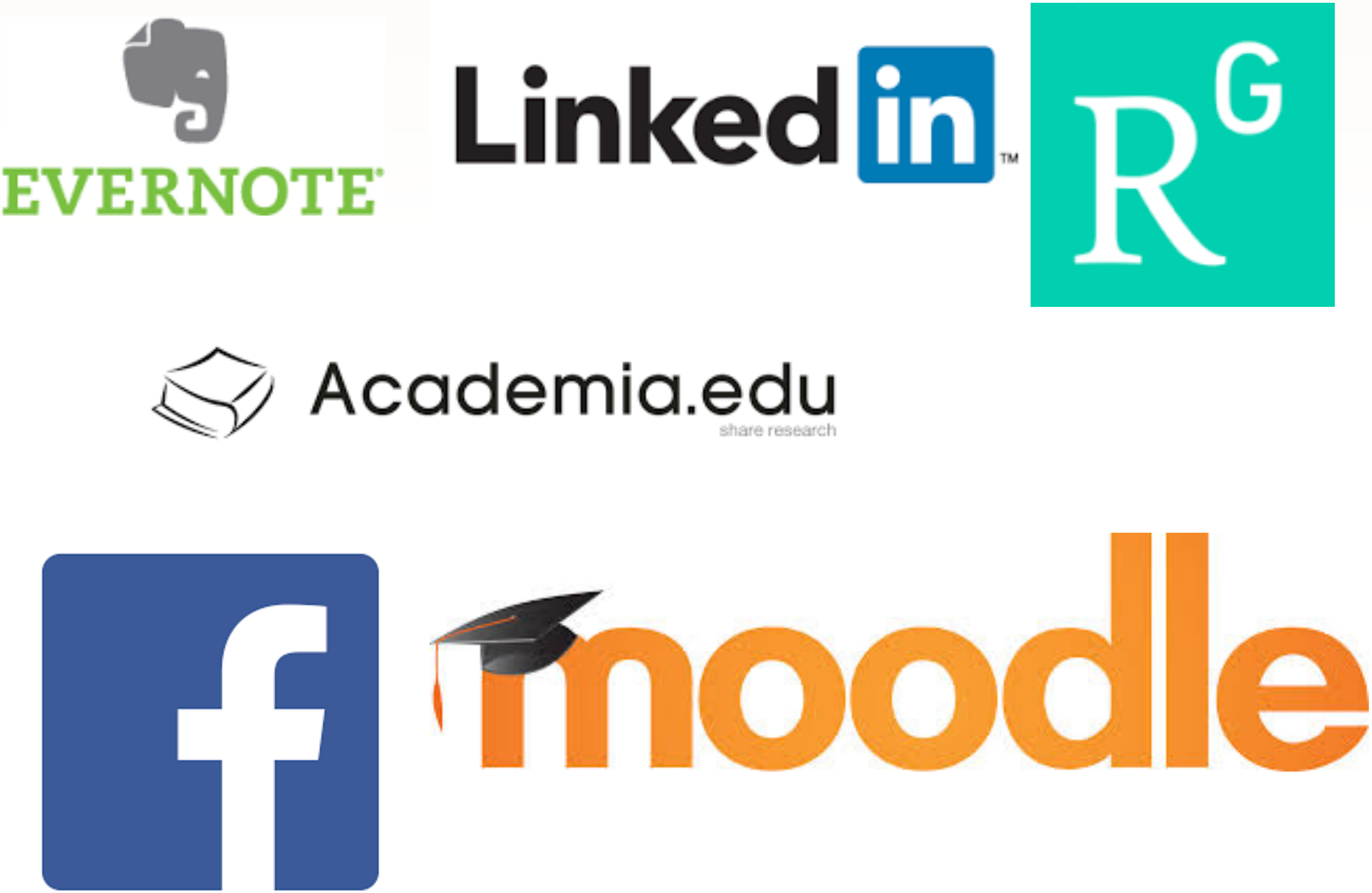
* Based on Avoiding data disasters course by Mark Dunning, CRUK-CI
http://bioinformatics-core-shared-training.github.io//avoid-data-disaster/

Warning!

- Every discipline is different
- These are general principles
- Application will vary according to your research

OSC

Where NOT to share your data



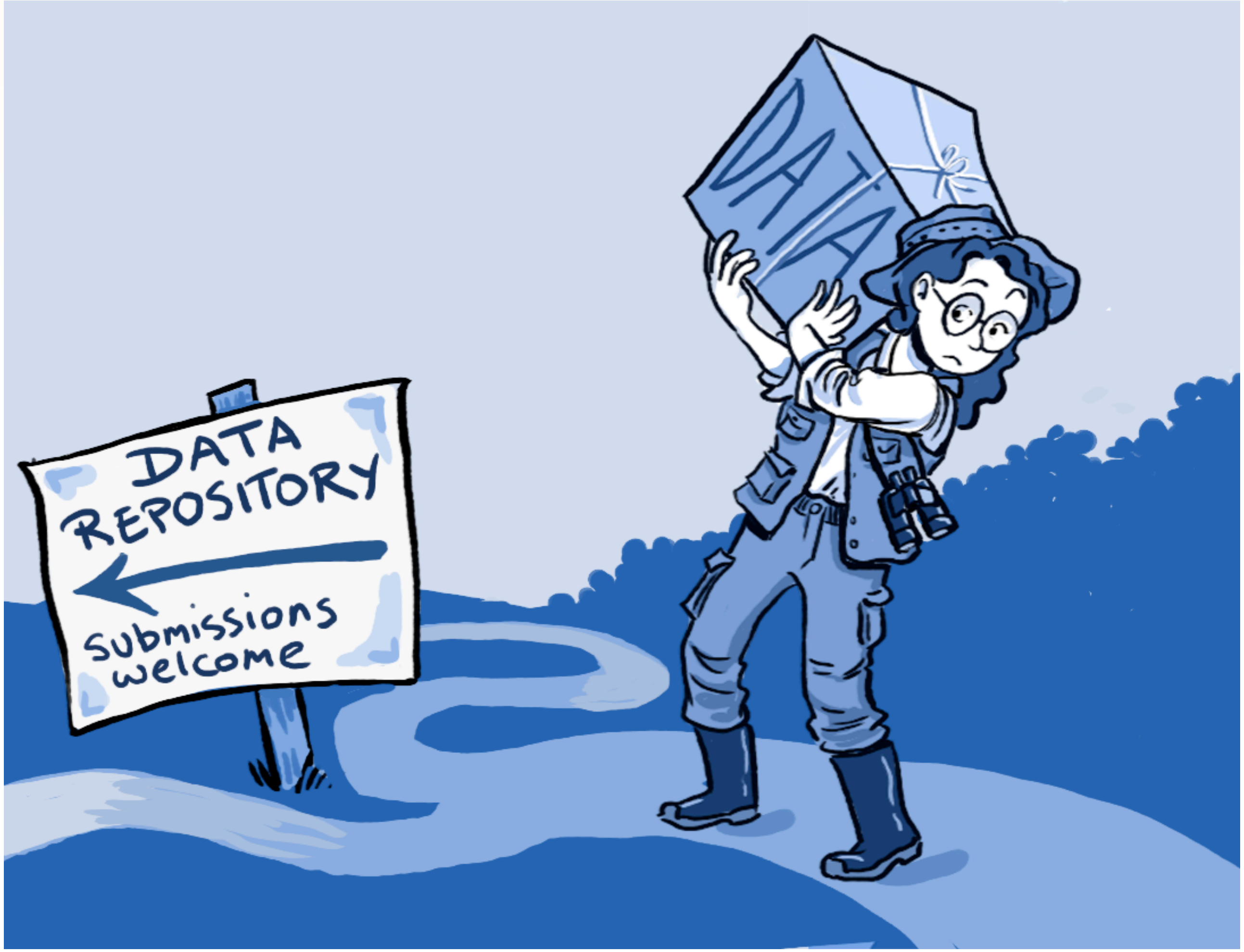
Where SHOULD you share your data?

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

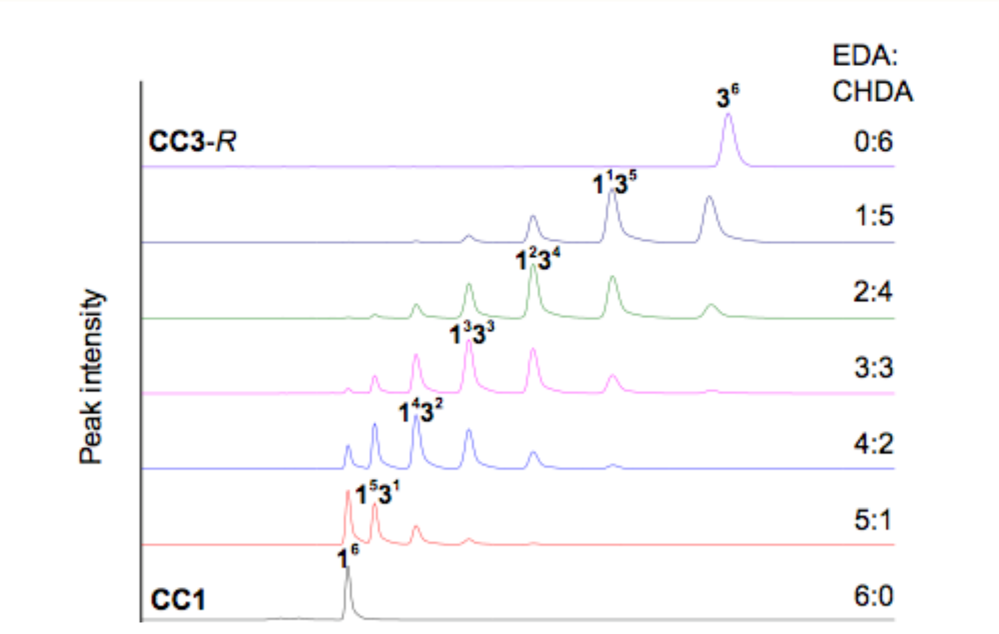
Apollo

GitHub → zenodo

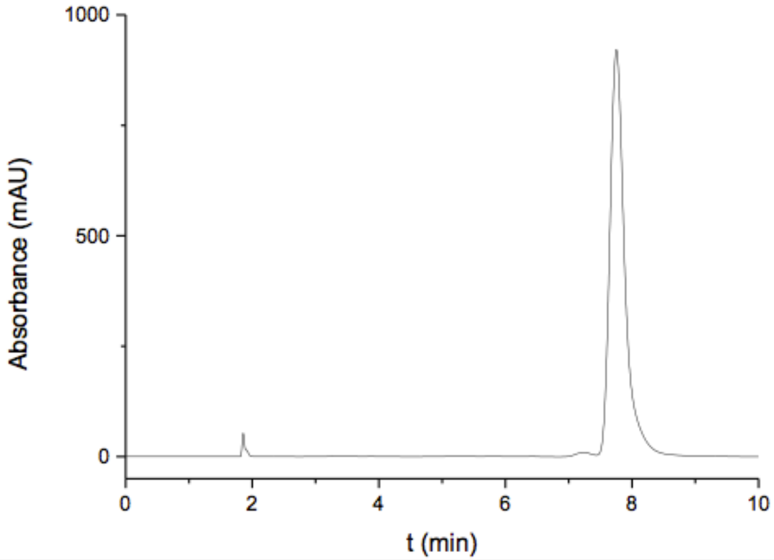


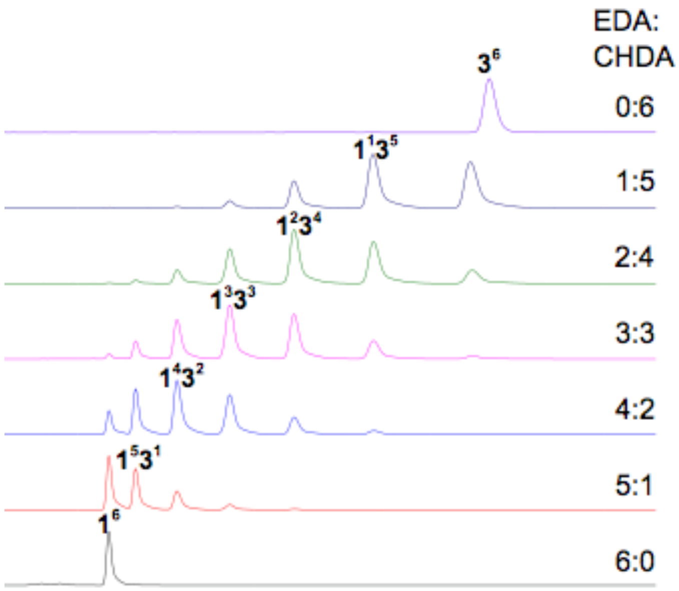
Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) - Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779. doi:10.1371/journal.pbio.1001779, CC BY 4.0,

What data should you include?



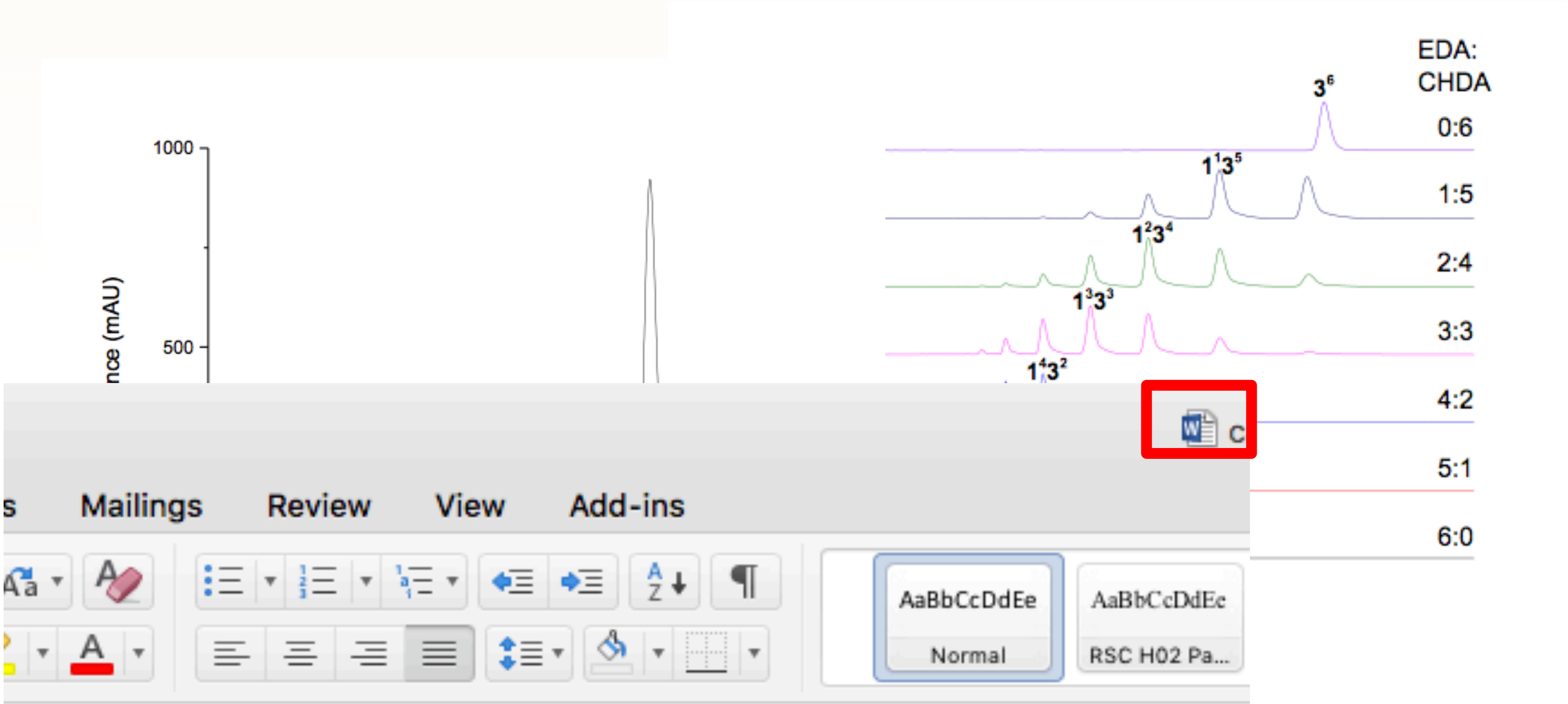
What data should you include?

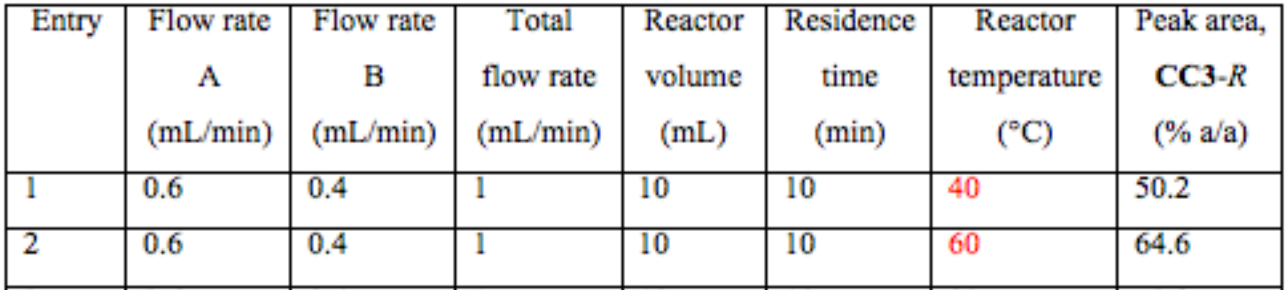




OSC

What data should you include?





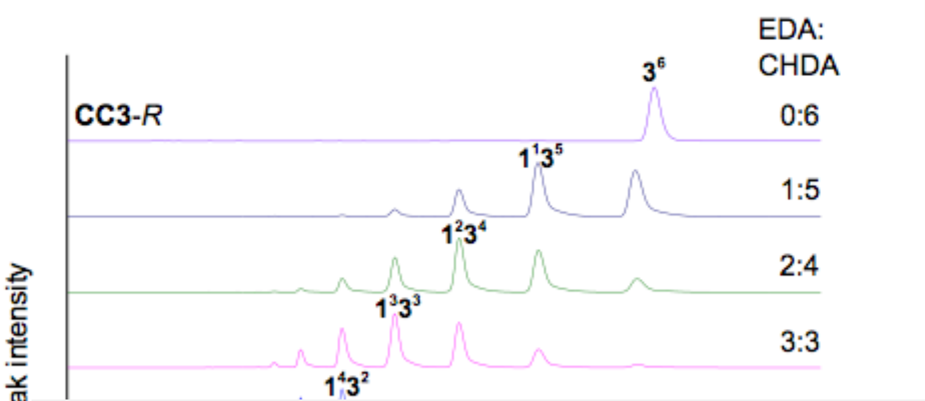
Entry	Flow rate A (mL/min)	Flow rate B (mL/min)	Total flow rate (mL/min)	Reactor volume (mL)	Residence time (min)	Reactor temperature (°C)	Peak area, **CC3-*R*** (% a/a)
1	0.6	0.4	1	10	10	40	50.2
2	0.6	0.4	1	10	10	60	64.6

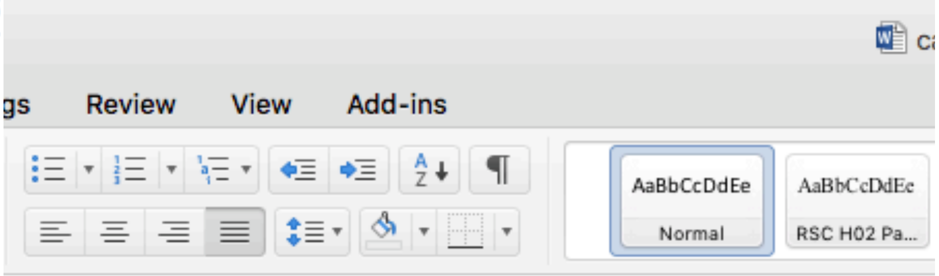
OSC

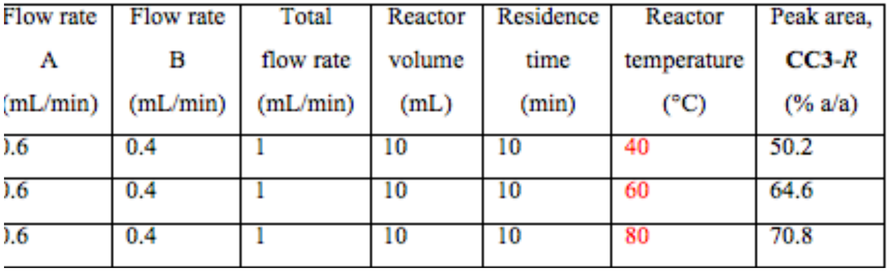
What data should you include?

More than your figures!

Data and code necessary to recreate your results



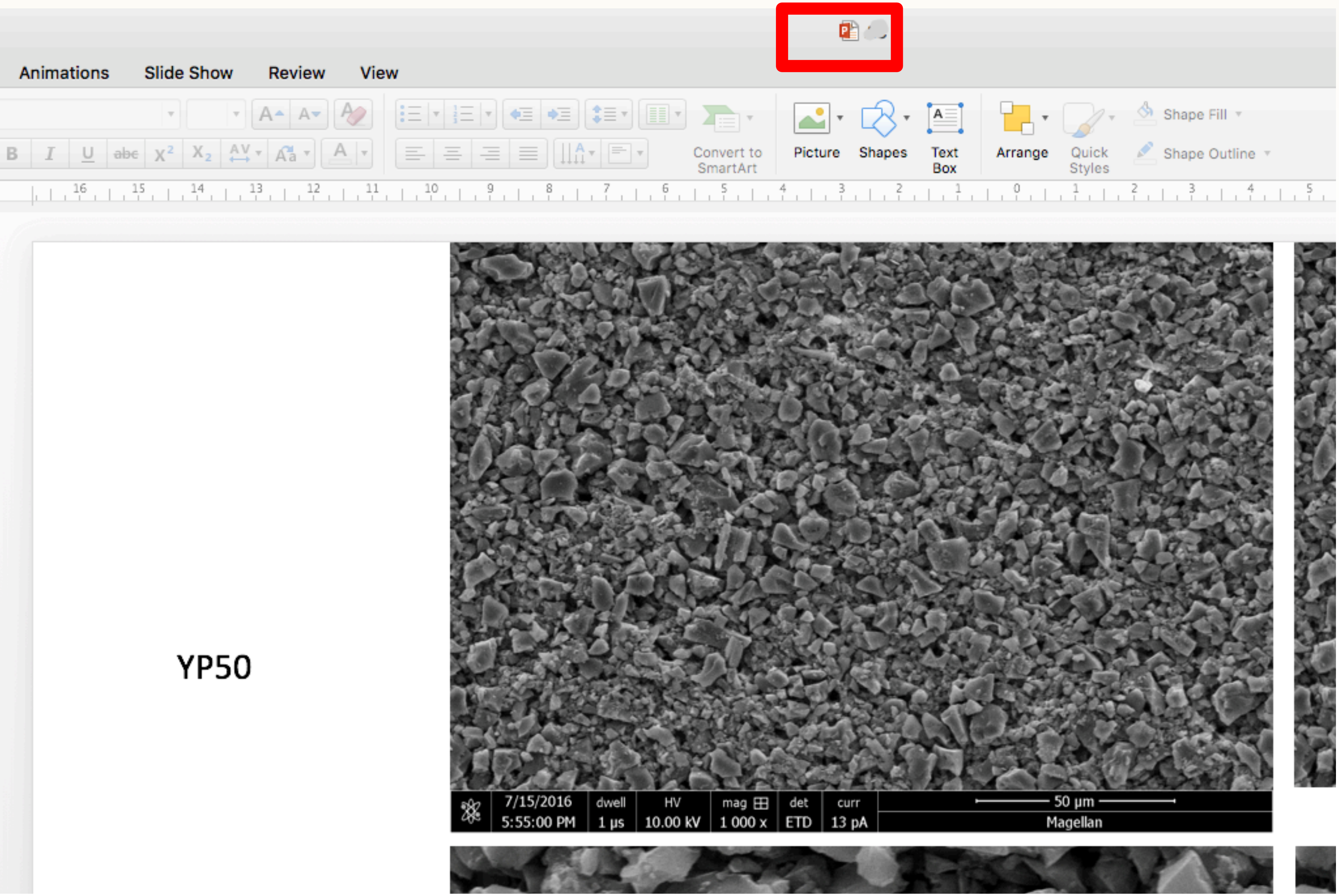




Flow rate A (mL/min)	Flow rate B (mL/min)	Total flow rate (mL/min)	Reactor volume (mL)	Residence time (min)	Reactor temperature (°C)	Peak area, CC3-*R* (% a/a)
).6	0.4	1	10	10	40	50.2
).6	0.4	1	10	10	60	64.6
).6	0.4	1	10	10	80	70.8

OSC

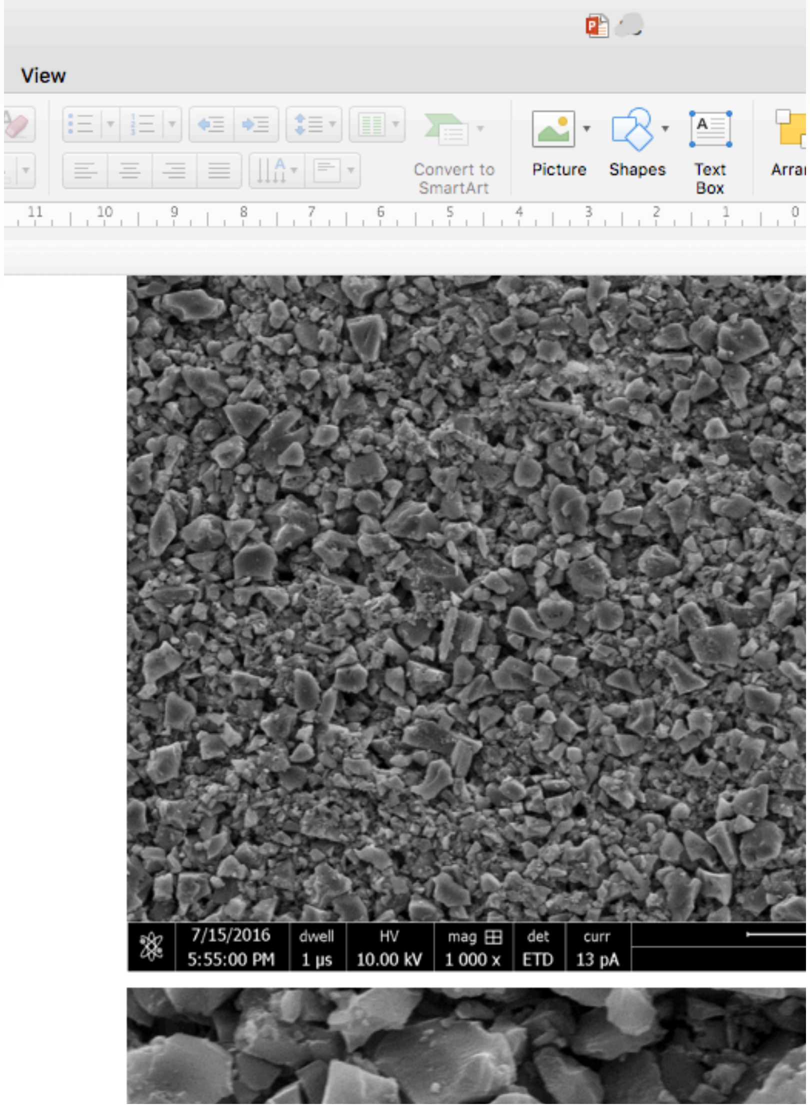
Powerpoint is for presentations NOT data!



Powerpoint is for presentations NOT data!

Instead:

Original image files
Appropriate formats
Annotations embedded in separate PDF/csv/txt file (README file)



7/15/2016 5:55:00 PM	dwell 1 µs	HV 10.00 kV	mag ⊞ 1 000 x	det ETD	curr 13 pA

Find your file format

Textual data = XML, TXT, HTML, PDF/A (Archival PDF)

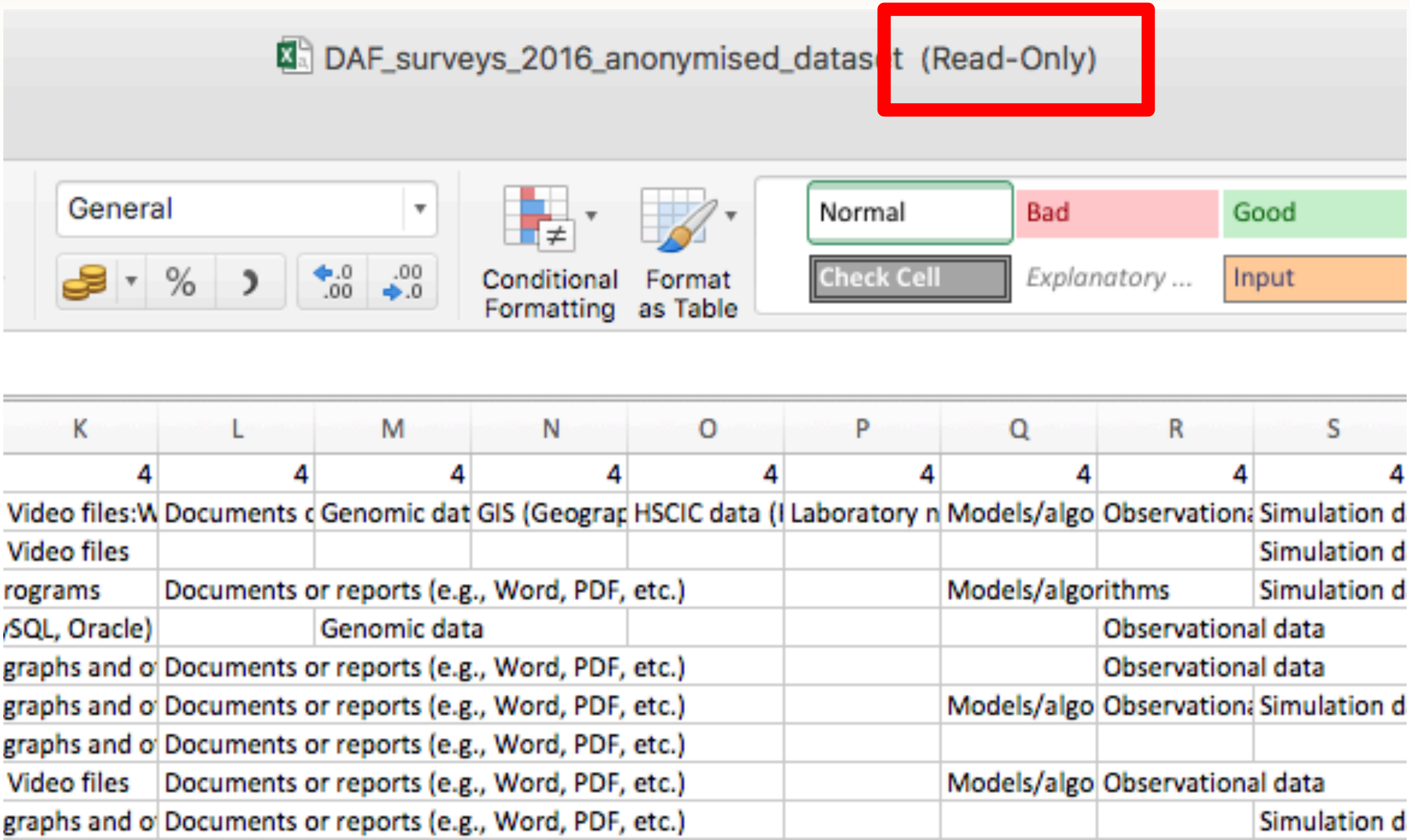
Tabular data (spreadsheets) = CSV

Databases = XML, CSV

Images = TIFF, PNG, JPEG*

Audio = FLAC, WAV, MP3

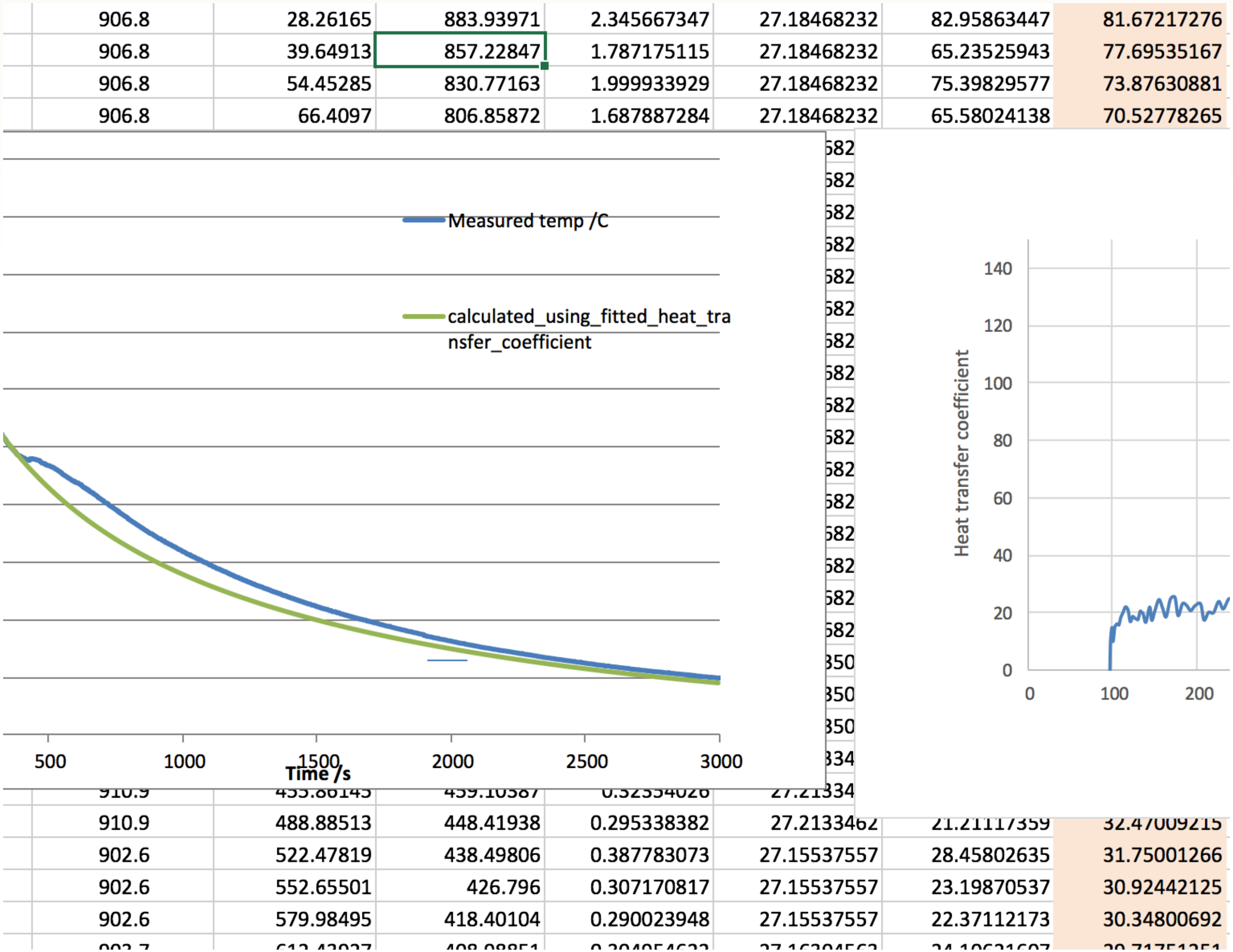
Think! Preservation vs access/re-use



DAF_surveys_2016_anonymised_dataset (Read-Only)

K	L	M	N	O	P	Q	R	S
4	4	4	4	4	4	4	4	4
Video files:W	Documents c	Genomic dat	GIS (Geograp	HSCIC data (I	Laboratory n	Models/algo	Observationa	Simulation d
Video files								Simulation d
rograms	Documents or reports (e.g., Word, PDF, etc.)					Models/algorithms		Simulation d
/SQL, Oracle)		Genomic data					Observational data	
graphs and o	Documents or reports (e.g., Word, PDF, etc.)						Observational data	
graphs and o	Documents or reports (e.g., Word, PDF, etc.)					Models/algo	Observationa	Simulation d
graphs and o	Documents or reports (e.g., Word, PDF, etc.)							
Video files	Documents or reports (e.g., Word, PDF, etc.)					Models/algo	Observational data	
graphs and o	Documents or reports (e.g., Word, PDF, etc.)							Simulation d

Messy spreadsheets are harder to re-use


Measured temp /C
calculated_using_fitted_heat_transfer_coefficient
500
1000
1500
2000
2500
3000
Time /s
Heat transfer coefficient
140
120
100
80
60
40
20
0
0
100
200

906.8	28.26165	883.93971	2.345667347	27.18468232	82.95863447	81.67217276
906.8	39.64913	857.22847	1.787175115	27.18468232	65.23525943	77.69535167
906.8	54.45285	830.77163	1.999933929	27.18468232	75.39829577	73.87630881
906.8	66.4097	806.85872	1.687887284	27.18468232	65.58024138	70.52778265
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				582		
				350		
				350		
				350		
				334		
910.9	455.86145	439.10387	0.32354026	27.21334		
910.9	488.88513	448.41938	0.295338382	27.2133462	21.21117359	32.47009215
902.6	522.47819	438.49806	0.387783073	27.15537557	28.45802635	31.75001266
902.6	552.65501	426.796	0.307170817	27.15537557	23.19870537	30.92442125
902.6	579.98495	418.40104	0.290023948	27.15537557	22.37112173	30.34800692
902.7	612.43027	408.98851	0.304054633	27.16304563	24.10621607	29.71751251

OSC

Keep your spreadsheets tidy

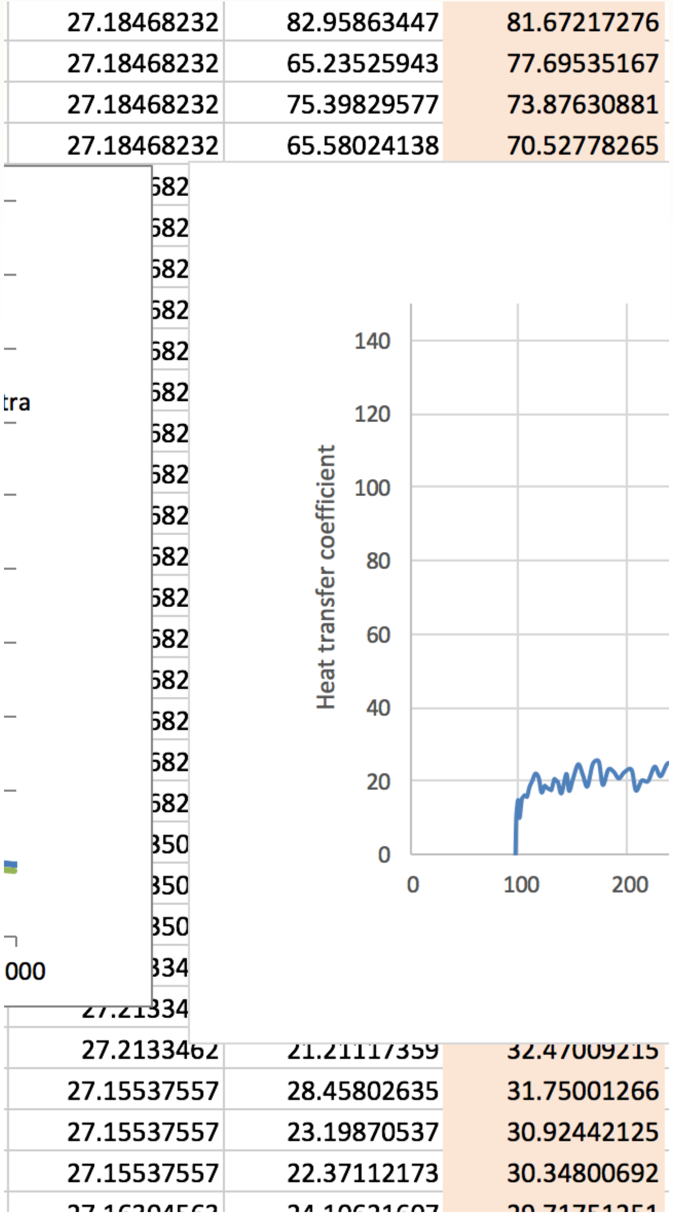
Graphs in separate sheet

No highlighting

No colours

No formulas*

*In your raw data



27.18468232	82.95863447	81.67217276
27.18468232	65.23525943	77.69535167
27.18468232	75.39829577	73.87630881
27.18468232	65.58024138	70.52778265
27.2133462	21.21117359	32.47009215
27.15537557	28.45802635	31.75001266
27.15537557	23.19870537	30.92442125
27.15537557	22.37112173	30.34800692

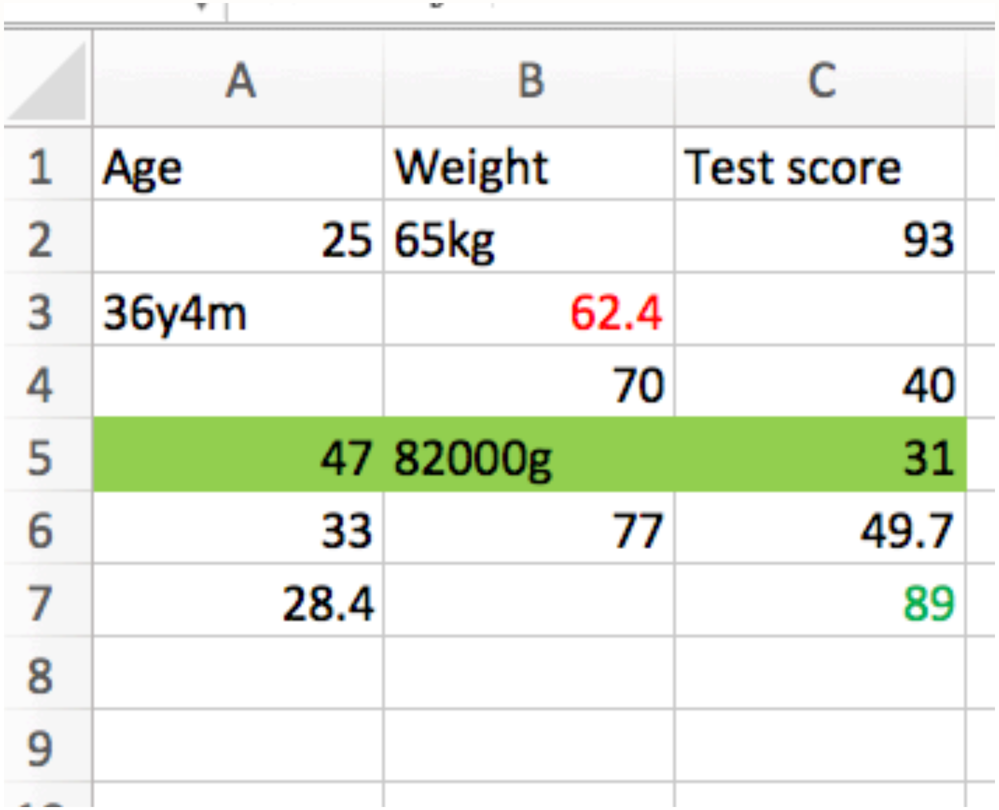
Keep your spreadsheets tidy

No blank cells

1 piece of data per cell

Keep units out of cells

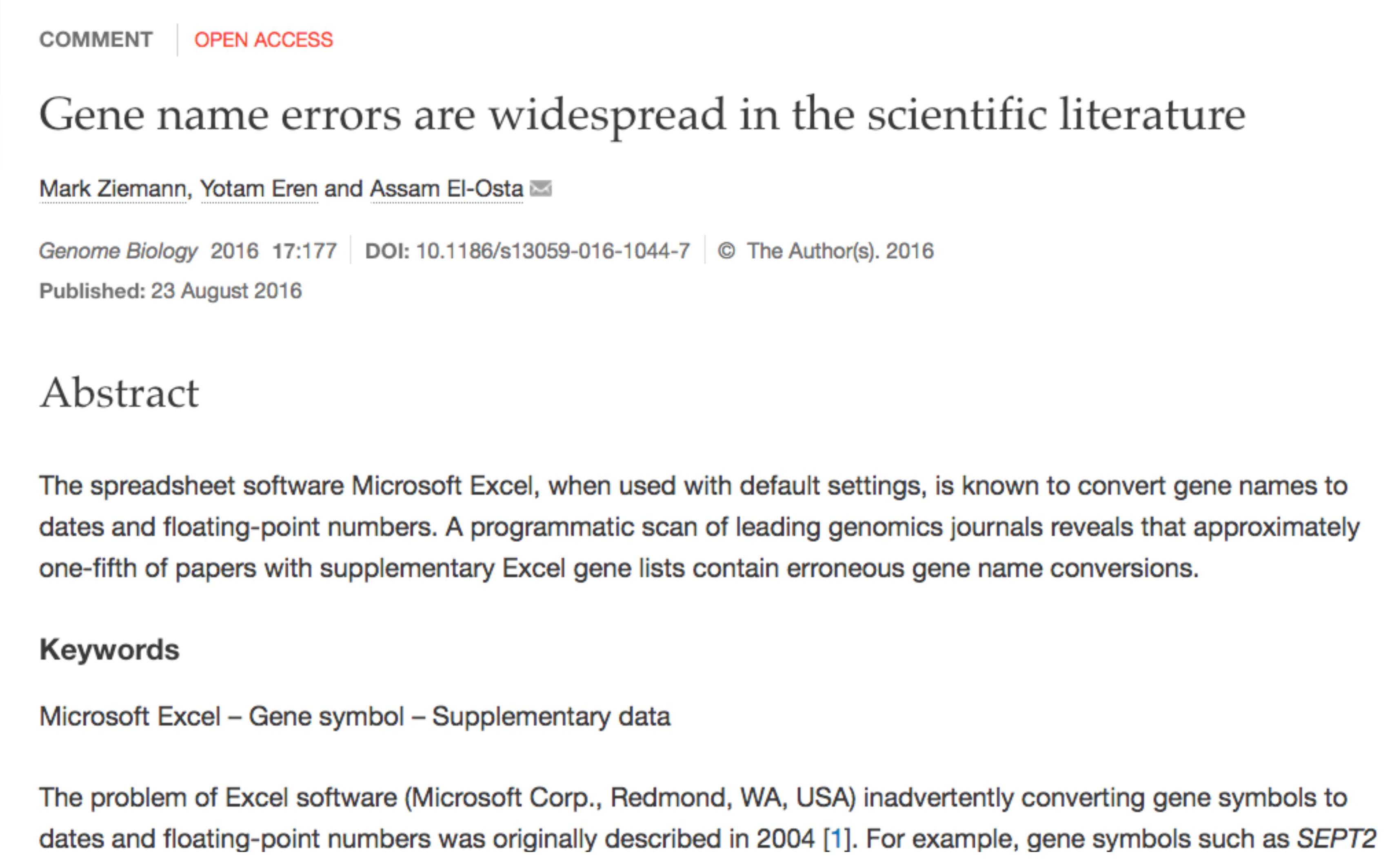
Use data validation



	A	B	C
1	Age	Weight	Test score
2	25	65kg	93
3	36y4m	62.4	
4		70	40
5	47	82000g	31
6	33	77	49.7
7	28.4		89
8			
9			

OSC

Keep your spreadsheets tidy



COMMENT | OPEN ACCESS

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta

Genome Biology 2016 17:177 | DOI: 10.1186/s13059-016-1044-7 | © The Author(s). 2016

Published: 23 August 2016

Abstract

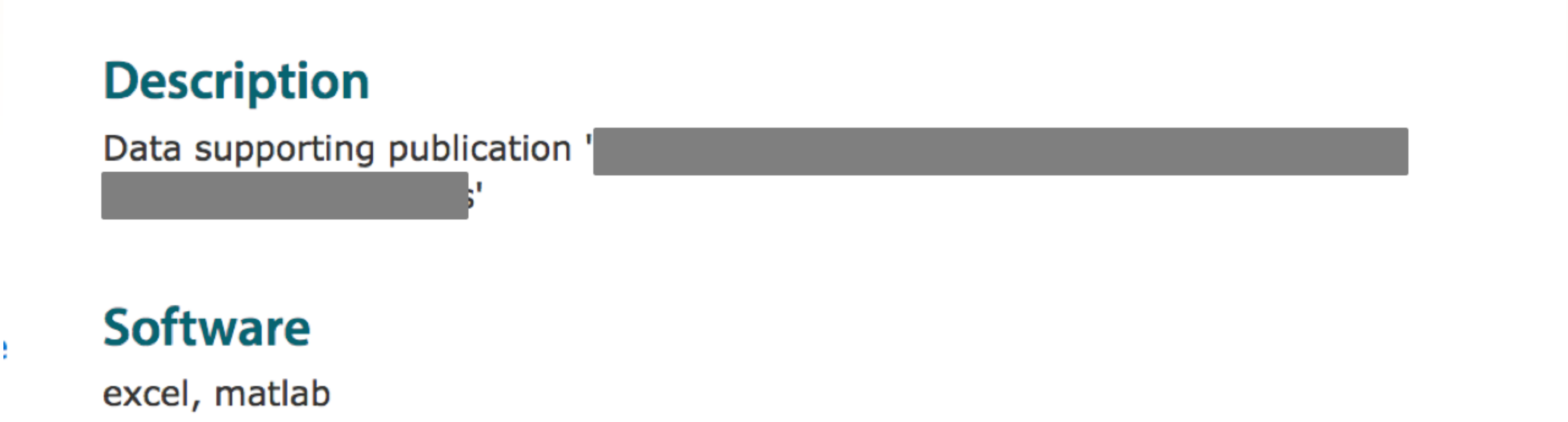
The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords

Microsoft Excel – Gene symbol – Supplementary data

The problem of Excel software (Microsoft Corp., Redmond, WA, USA) inadvertently converting gene symbols to dates and floating-point numbers was originally described in 2004 [1]. For example, gene symbols such as *SEPT2*

How NOT to describe your data



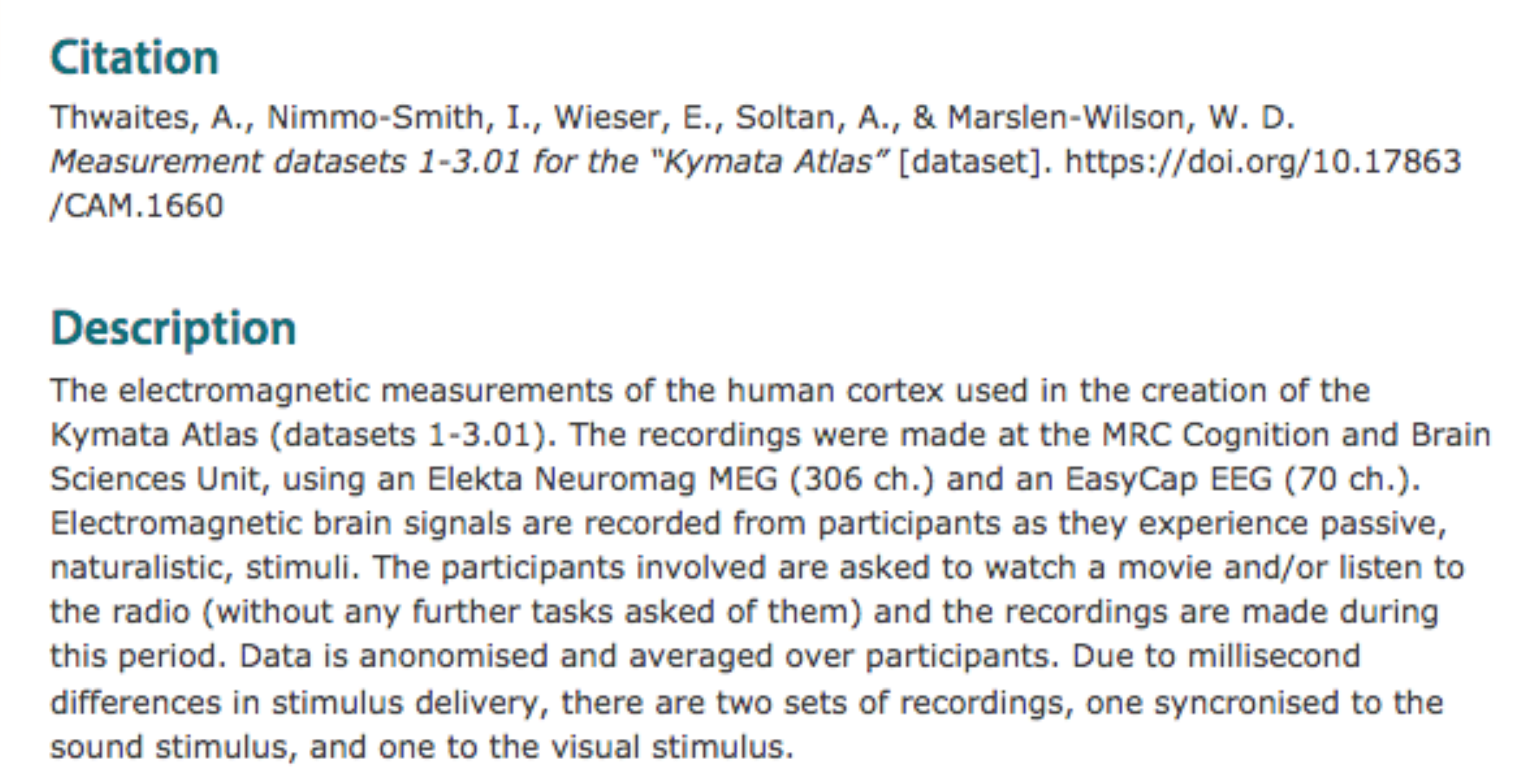
Description

Data supporting publication '[redacted]'

Software

excel, matlab

How you SHOULD describe your data



Citation

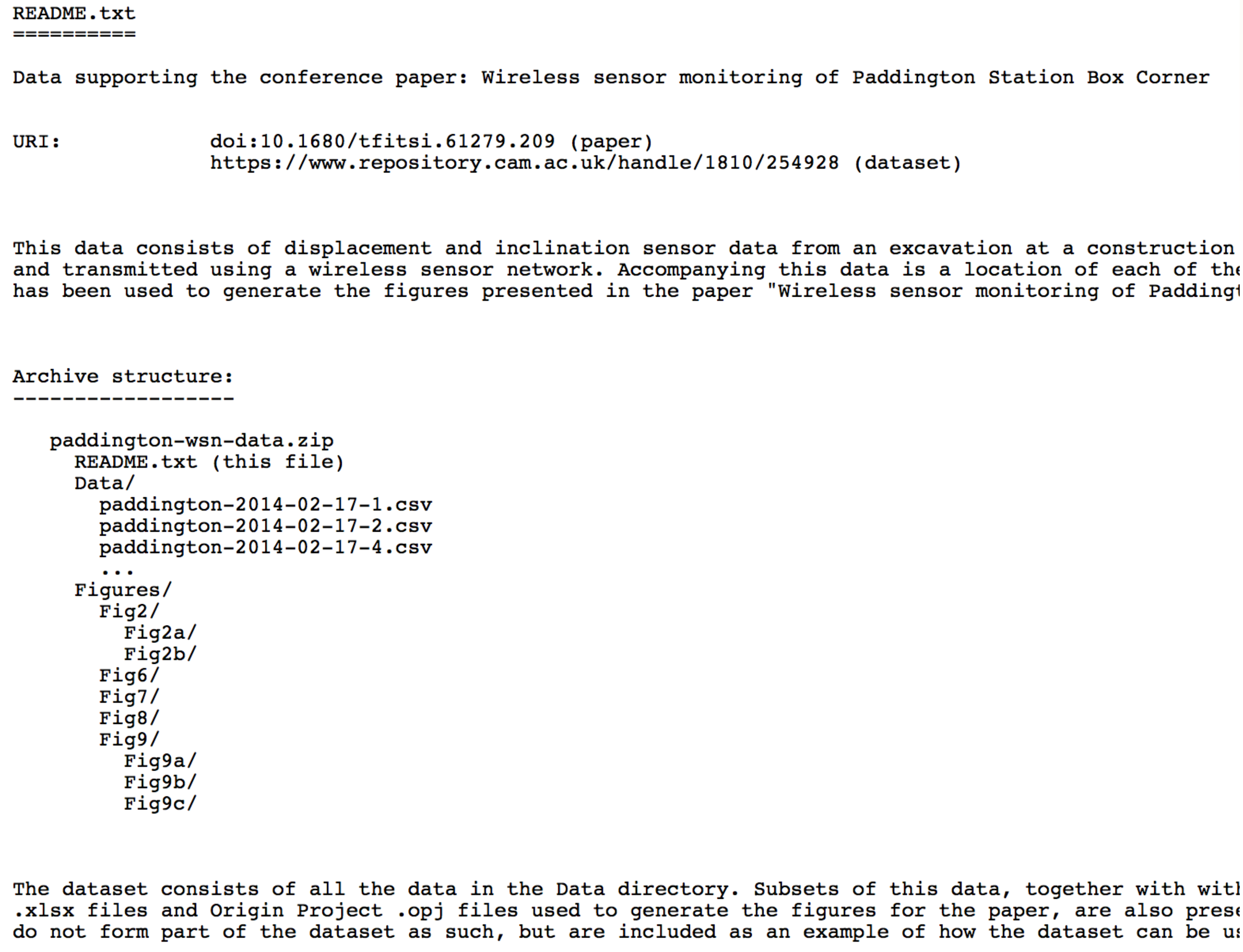
Thwaites, A., Nimmo-Smith, I., Wieser, E., Soltan, A., & Marslen-Wilson, W. D. *Measurement datasets 1-3.01 for the "Kymata Atlas"* [dataset]. https://doi.org/10.17863/CAM.1660

Description

The electromagnetic measurements of the human cortex used in the creation of the Kymata Atlas (datasets 1-3.01). The recordings were made at the MRC Cognition and Brain Sciences Unit, using an Elekta Neuromag MEG (306 ch.) and an EasyCap EEG (70 ch.). Electromagnetic brain signals are recorded from participants as they experience passive, naturalistic, stimuli. The participants involved are asked to watch a movie and/or listen to the radio (without any further tasks asked of them) and the recordings are made during this period. Data is anonomised and averaged over participants. Due to millisecond differences in stimulus delivery, there are two sets of recordings, one syncronised to the sound stimulus, and one to the visual stimulus.

OSC

Remember to document your data



```
README.txt
==========

Data supporting the conference paper: Wireless sensor monitoring of Paddington Station Box Corner


URI:            doi:10.1680/tfitsi.61279.209 (paper)
                https://www.repository.cam.ac.uk/handle/1810/254928 (dataset)



This data consists of displacement and inclination sensor data from an excavation at a construction
and transmitted using a wireless sensor network. Accompanying this data is a location of each of the
has been used to generate the figures presented in the paper "Wireless sensor monitoring of Paddingt



Archive structure:
------------------

   paddington-wsn-data.zip
     README.txt (this file)
     Data/
       paddington-2014-02-17-1.csv
       paddington-2014-02-17-2.csv
       paddington-2014-02-17-4.csv
       ...
     Figures/
       Fig2/
         Fig2a/
         Fig2b/
       Fig6/
       Fig7/
       Fig8/
       Fig9/
         Fig9a/
         Fig9b/
         Fig9c/


The dataset consists of all the data in the Data directory. Subsets of this data, together with with
.xlsx files and Origin Project .opj files used to generate the figures for the paper, are also prese
do not form part of the dataset as such, but are included as an example of how the dataset can be us
```

Choose a repository.
Choose open file formats.
Choose sharing more than your figures.
Choose a tidy spreadsheet.
Choose to describe your data.
Choose decent documentation so your research is reproducible.

CHOOSE DATA SHARING