R: A Hitchhikers Guide to Reproducible Research

- My favourite mistake

Brendan Palmer,
Clinical Research Facility - Cork &
School of Public Health
@B_A_Palmer

CRF-C
HRB Clinical Research Facility Cork

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of Public Health

Fundamental problem

I'm not in the office at the moment. Send any work to be translated

Beware of default settings

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT Open Access

Gene name errors are widespread in the scientific literature

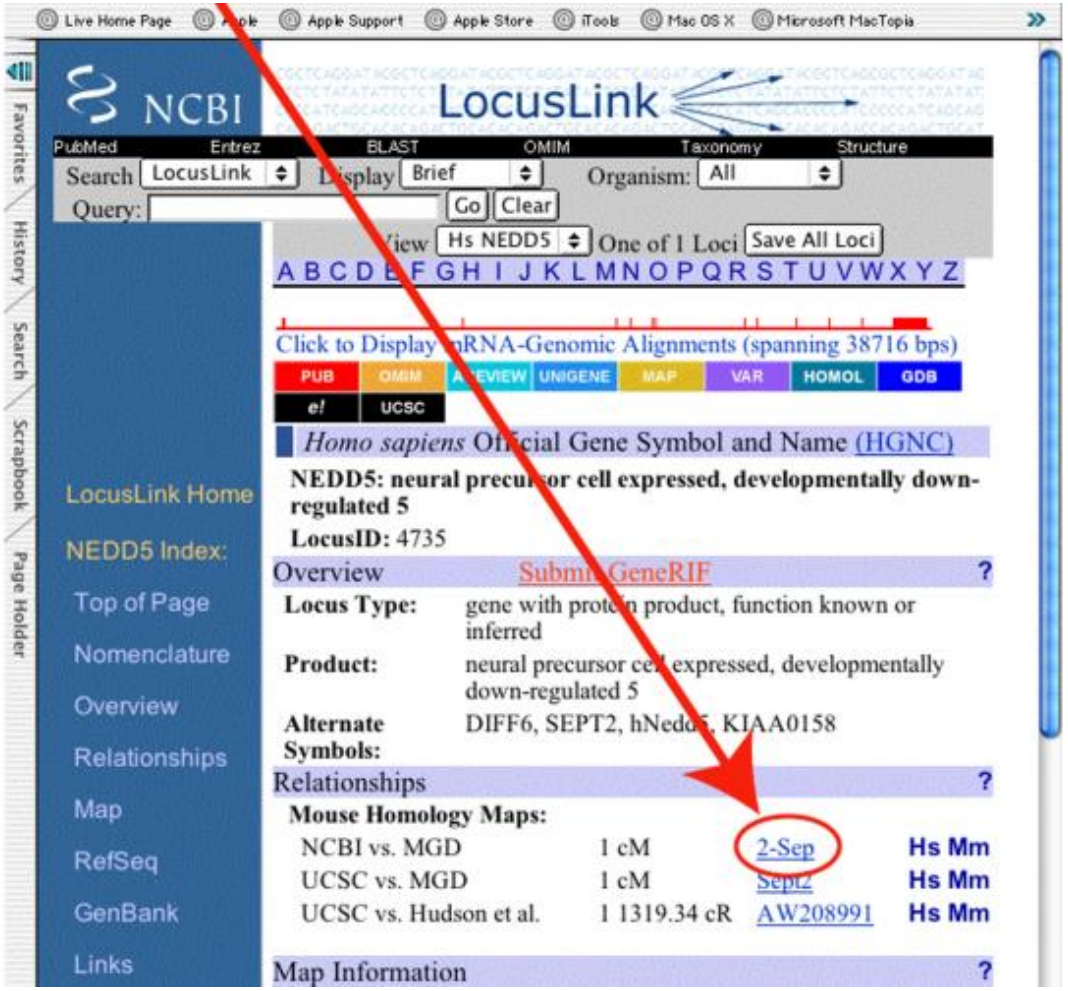
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

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.

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-



Take small steps to big changes

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

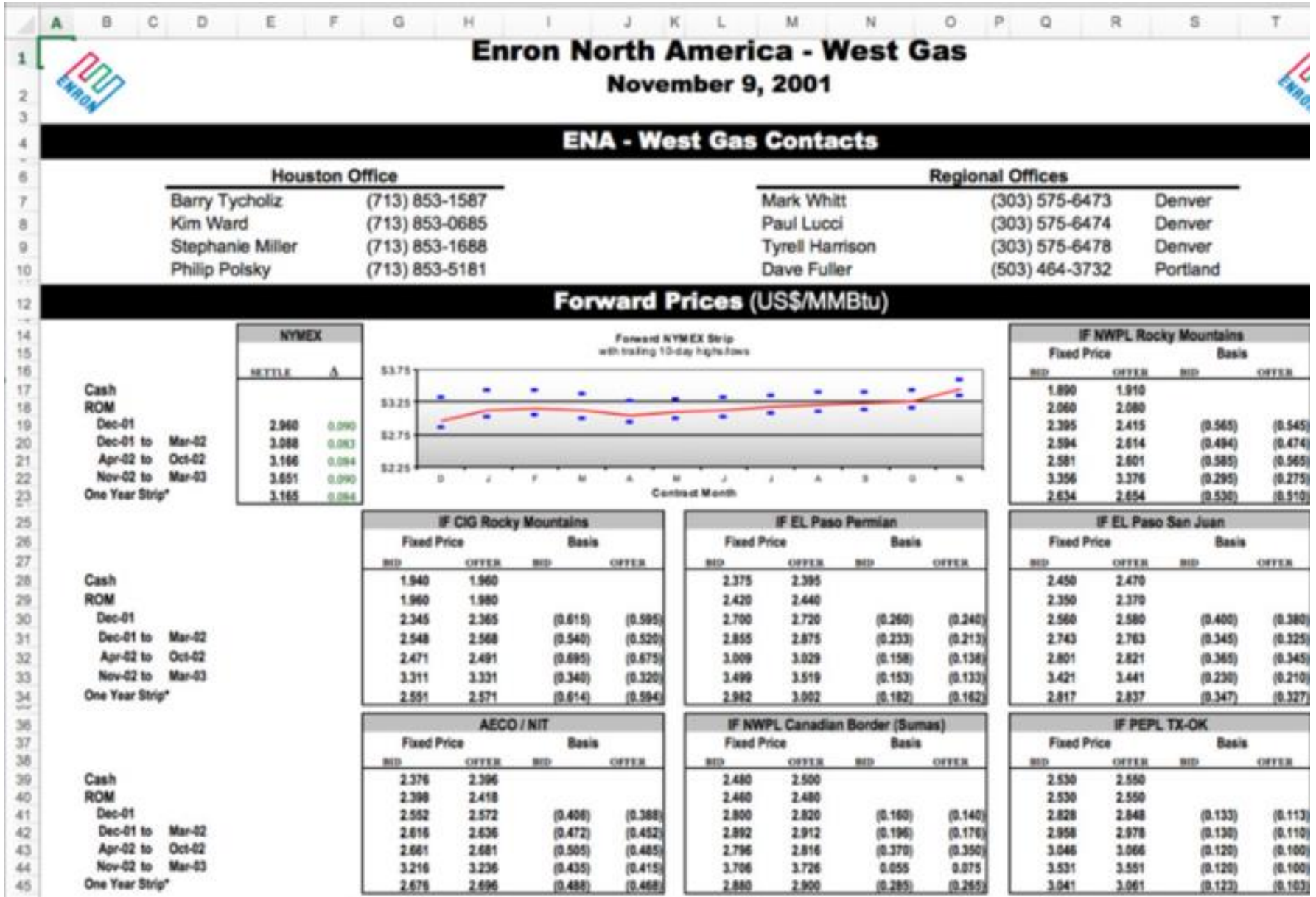
ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

This is a big problem



Enron North America - West Gas

November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX SETTLE	NYMEX Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084

Forward NYMEX Strip
with trailing 10-day highs/lows

$3.75
$3.25
$2.75
$2.25

D J F M A M J J A S O N

Contract Month

	IF NWPL Rocky Mountains			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.890	1.910		
ROM	2.060	2.080		
Dec-01	2.395	2.415	(0.565)	(0.545)
Dec-01 to Mar-02	2.594	2.614	(0.494)	(0.474)
Apr-02 to Oct-02	2.581	2.601	(0.585)	(0.565)
Nov-02 to Mar-03	3.356	3.376	(0.295)	(0.275)
One Year Strip*	2.634	2.654	(0.530)	(0.510)

	IF CIG Rocky Mountains			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

	IF EL Paso Permian			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

	IF EL Paso San Juan			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

	AECO / NIT			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

	IF NWPL Canadian Border (Sumas)			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

	IF PEPL TX-OK			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp	nofilter	2	1.1805	0.42	cos	2019/04/01	Darren Dahly		
4	3	0	ptp	nofilter	3	1.0345	0.62	cos	2019/04/01	Darren Dahly		
5	4	0	ptp	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
6	1	0	my	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
7	2	0	my	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
8	3	0	my	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
9	4	0	my	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
10	1	0	ca	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
11	2	0	ca	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
12	3	0	ca	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
13	4	0	ca	nofilter	4	1.094	0.63	cos	2019/04/01	Darren Dahly		
14	5	1	ptp	filter	1	0.87	0.76	cos	2019/04/08	Darren Dahly		
15	6	1	ptp	filter	2	0.847	0.95	cos	2019/04/08	Darren Dahly		
16	7	1	ptp	filter	3	1.022	0.95	cos	2019/04/08	Darren Dahly		
17	8	1	ptp	filter	4	0.916	0.95	cos	2019/04/08	Darren Dahly		
18	9	1	my	filter	1	1.119	1.55	cos	2019/04/08	Darren Dahly		
19	10	1	my	filter	2	0.845	3.16	cos	2019/04/08	Darren Dahly		
20	11	1	my	filter	3	1.299	4.9	cos	2019/04/08	Brendan Palmer		
21	12	1	my	filter	4	1.149	5.5	cos	2019/04/08	Brendan Palmer		
22	13	1	ca	filter	1	0.716	5.5	cos	2019/04/08	Brendan Palmer		
23	14	1	ca	filter	2	0.881	7.94	cos	2019/04/08	Brendan Palmer		
24	15	1	ca	filter	3	0.586	8.71	cos	2019/04/08	Brendan Palmer		
25	16	1	ca	filter	4	0.561	8.71	cos	2019/04/08	Brendan Palmer		
26	17	2	ptp	filter	1	0	14.45	cos	2019/04/15	Brendan Palmer		
27	18	2	ptp	filter	2	1.006	2.14	cos	2019/04/15	Brendan Palmer		
28	19	2	ptp	filter	3	1.236	1.86	cos	2019/04/15	Brendan Palmer		
29	20	2	ptp	filter	4	1.206	1.2	cos	2019/04/15	Brendan Palmer		
30	21	2	my	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer		

data | dictionary | values

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp									
4	3	0	ptp									
5	4	0	ptp									
6	1	0	my									
7	2	0	my									
8	3	0	my									
9	4	0	my									
10	1	0	ca									
11	2	0	ca									
12	3	0	ca									
13	4	0	ca									
14	5	1	ptp									
15	6	1	ptp									
16	7	1	ptp									
17	8	1	ptp									
18	9	1	my									
19	10	1	my									
20	11	1	my									
21	12	1	my									
22	13	1	ca									
23	14	1	ca									
24	15	1	ca									
25	16	1	ca									
26	17	2	ptp									
27	18	2	ptp									
28	19	2	ptp									
29	20	2	ptp									
30	21	2	my									

data | dictionary

	A	B	C	D	E	F
1	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer	1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
4	filter_name	character	NA	my	NA	3 filter types; 'ptp' = polytunnel plastic blocks all UV light
5	treatment	character	NA	filter	NA	Presence or absence of a filter at the time of sampling
6	replicate_no	numeric	integer	1	NA	The number of replicates in each treatement
7	flavonoids	numeric	double	0.3421	parts per million (ppm)	Leaf disc taken from the tip of the most mature leaf at th
8	biomass	numeric	double		gram (g)	Above ground biomass on the day of harvest
9	variety	character	NA	cos	NA	3 commerical varieties of red lettuce used; 'cos' = Cos Di
10	date	date	YYYY/MM/DD	2019/06/28	ISO 8601	Experiment date
11	investigator	character	Firstname Lastname	Aoife Coffey	NA	Primary researcher who performed the experiment
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						

data | dictionary | values

Less stress, more success

	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly
3	2	0	ptp							
4	3	0	ptp							
5	4	0	ptp							
6	1	0	my							
7	2	0	my							
8	3	0	my							
9	4	0	my							
10	1	0	ca							
11	2	0	ca							
12	3	0	ca							
13	4	0	ca							
14	5	1	ptp							
15	6	1	ptp							
16	7	1	ptp							
17	8	1	ptp							
18	9	1	my							
19	10	1	my							
20	11	1	my							
21	12	1	my							
22	13	1	ca							
23	14	1	ca							
24	15	1	ca							
25	16	1	ca							
26	17	2	ptp							
27	18	2	ptp							
28	19	2	ptp							
29	20	2	ptp							
30	21	2	my							

data | dictionary

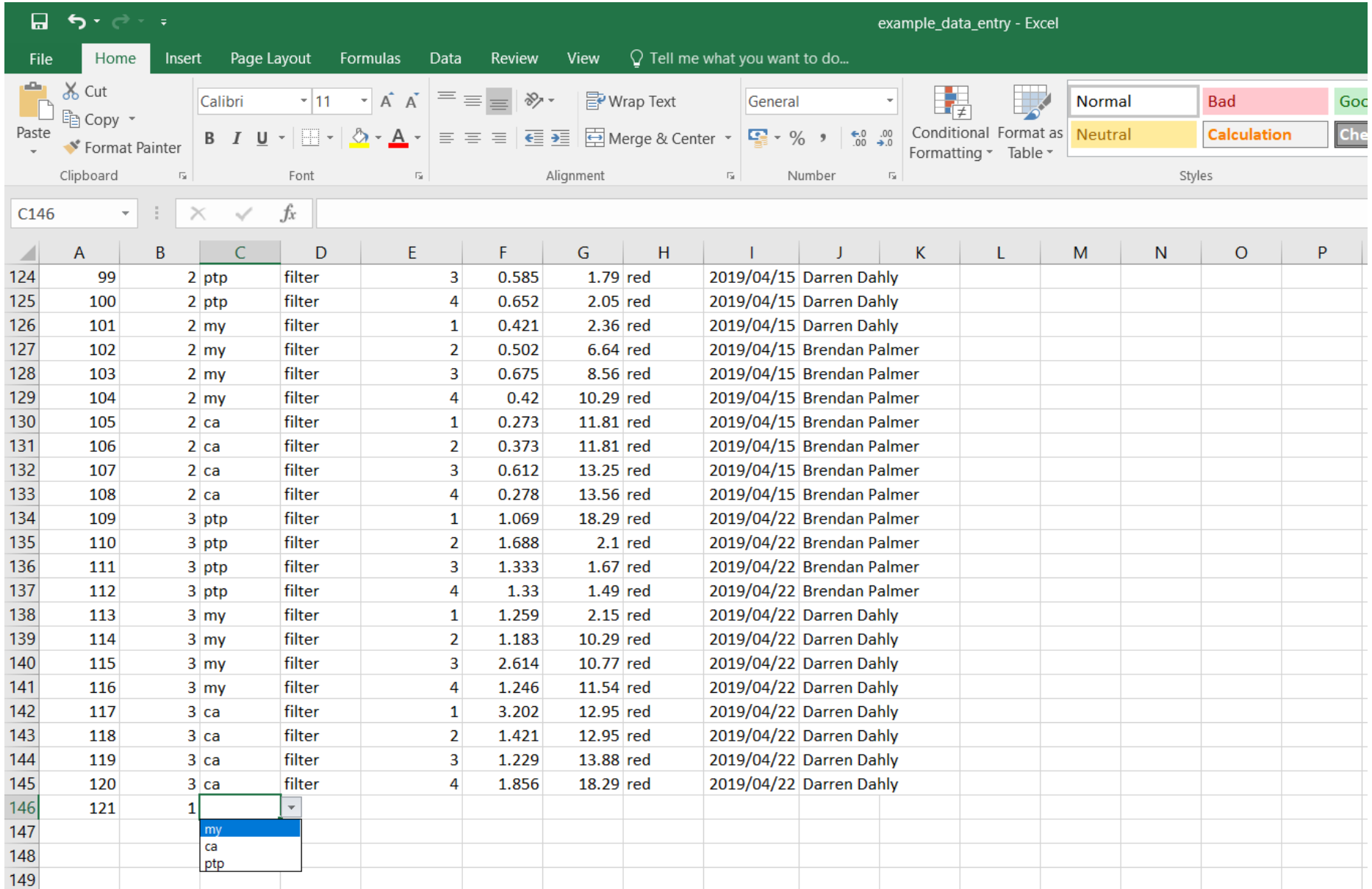
	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer			
4	filter_name	character	NA	n		
5	treatment	character	NA	fi		
6	replicate_no	numeric	integer			
7	flavonoids	numeric	double			
8	biomass	numeric	double			
9	variety	character	NA	c		
10	date	date	YYYY/MM/DD			
11	investigator	character	Firstname Lastname	A		

data | dictionary | values

	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
2		0	my	filter	1			cos		Brendan Palmer
3		1	ca	no_filter	2			sky		Darren Dahly
4		2	ptp		3			red		
5		3			4					

data | dictionary | values

Less stress, more success



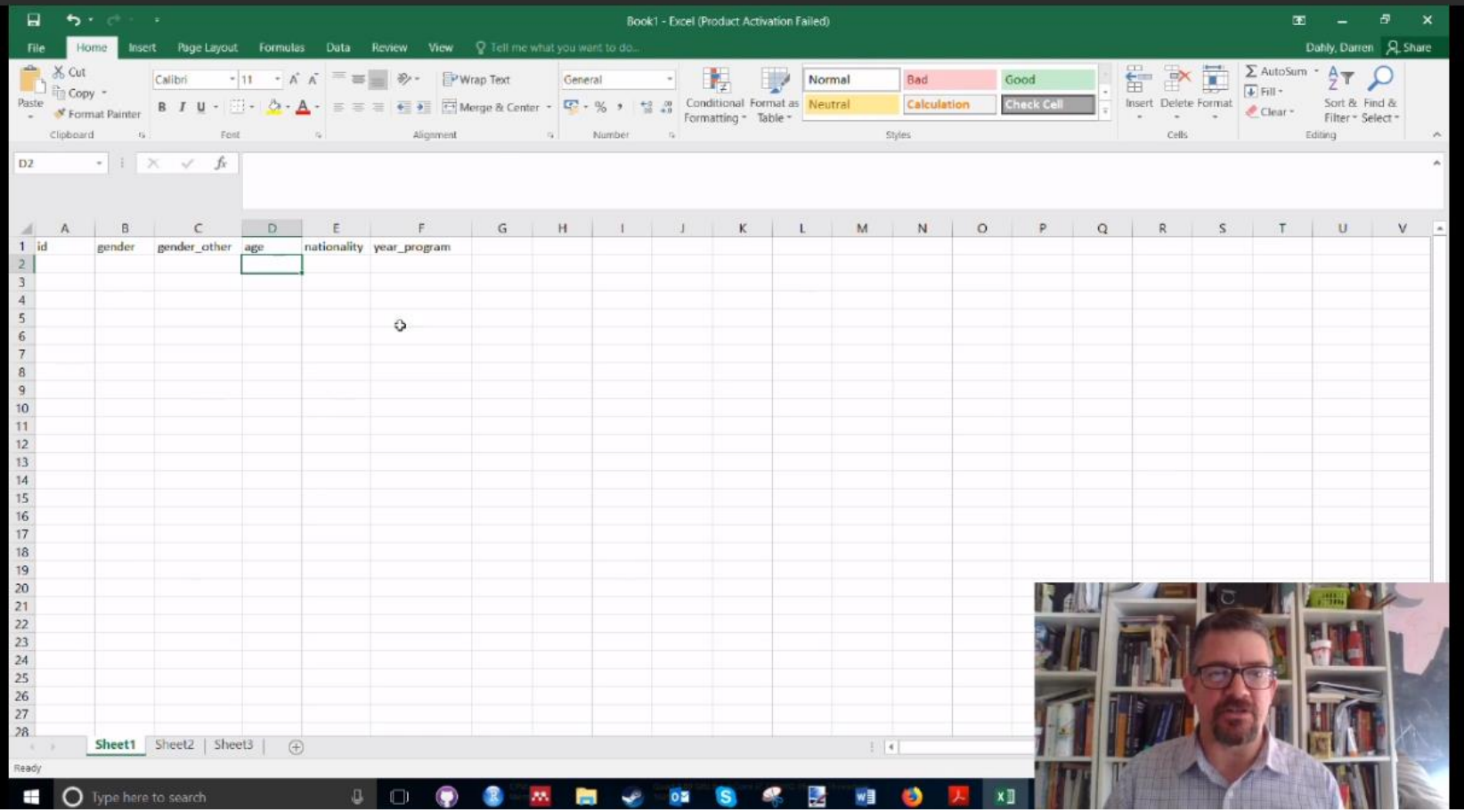
	A	B	C	D	E	F	G	H	I	J
124	99	2	ptp	filter	3	0.585	1.79	red	2019/04/15	Darren Dahly
125	100	2	ptp	filter	4	0.652	2.05	red	2019/04/15	Darren Dahly
126	101	2	my	filter	1	0.421	2.36	red	2019/04/15	Darren Dahly
127	102	2	my	filter	2	0.502	6.64	red	2019/04/15	Brendan Palmer
128	103	2	my	filter	3	0.675	8.56	red	2019/04/15	Brendan Palmer
129	104	2	my	filter	4	0.42	10.29	red	2019/04/15	Brendan Palmer
130	105	2	ca	filter	1	0.273	11.81	red	2019/04/15	Brendan Palmer
131	106	2	ca	filter	2	0.373	11.81	red	2019/04/15	Brendan Palmer
132	107	2	ca	filter	3	0.612	13.25	red	2019/04/15	Brendan Palmer
133	108	2	ca	filter	4	0.278	13.56	red	2019/04/15	Brendan Palmer
134	109	3	ptp	filter	1	1.069	18.29	red	2019/04/22	Brendan Palmer
135	110	3	ptp	filter	2	1.688	2.1	red	2019/04/22	Brendan Palmer
136	111	3	ptp	filter	3	1.333	1.67	red	2019/04/22	Brendan Palmer
137	112	3	ptp	filter	4	1.33	1.49	red	2019/04/22	Brendan Palmer
138	113	3	my	filter	1	1.259	2.15	red	2019/04/22	Darren Dahly
139	114	3	my	filter	2	1.183	10.29	red	2019/04/22	Darren Dahly
140	115	3	my	filter	3	2.614	10.77	red	2019/04/22	Darren Dahly
141	116	3	my	filter	4	1.246	11.54	red	2019/04/22	Darren Dahly
142	117	3	ca	filter	1	3.202	12.95	red	2019/04/22	Darren Dahly
143	118	3	ca	filter	2	1.421	12.95	red	2019/04/22	Darren Dahly
144	119	3	ca	filter	3	1.229	13.88	red	2019/04/22	Darren Dahly
145	120	3	ca	filter	4	1.856	18.29	red	2019/04/22	Darren Dahly
146	121	1								
147										
148										
149										

Step by step guide

https://www.youtube.com/watch?v=Ry2xjTBtNFE

YouTube IE

Search



The movement towards FAIR data

F + **A** + **I** = **R**




Findable **Accessible** **Interoperable**

Reusable
(Replicable,
Reproducible)

www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

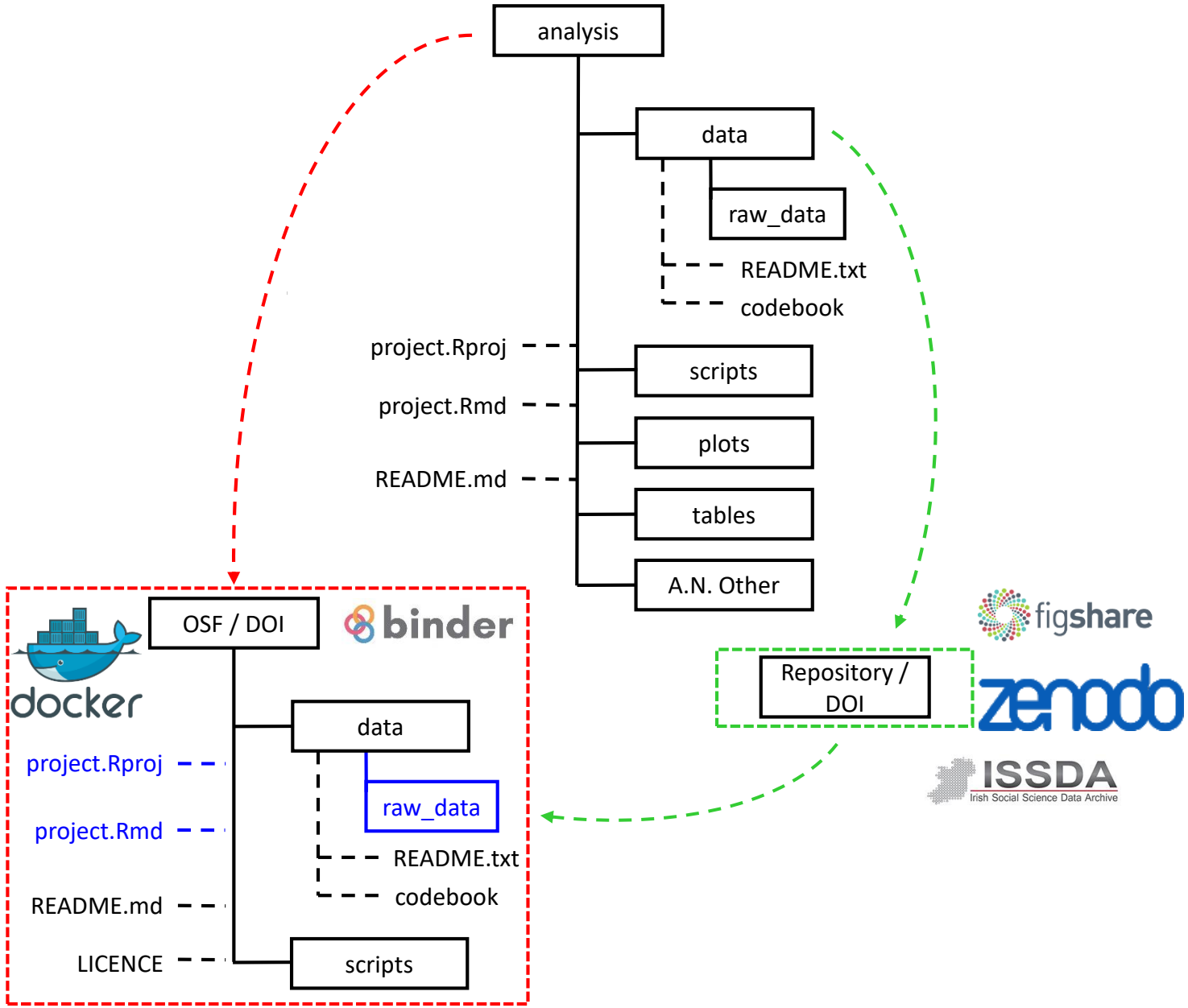
OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

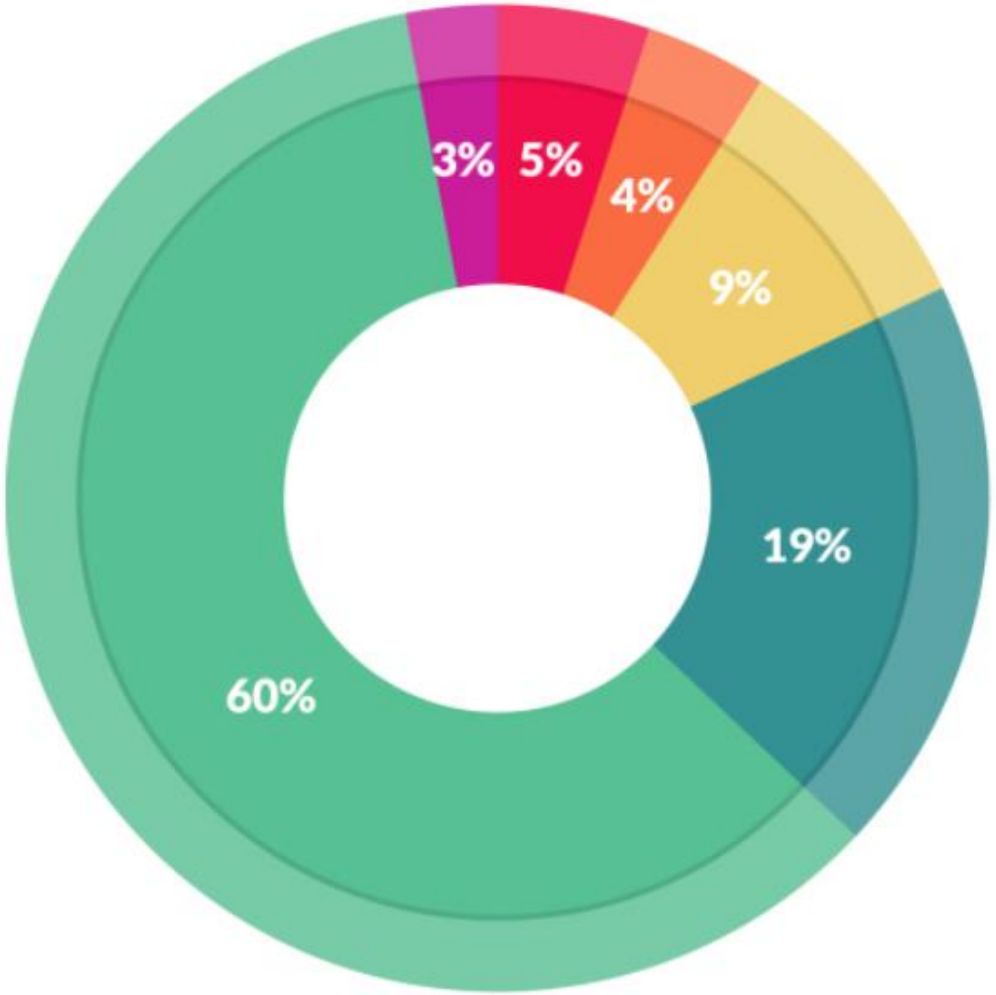
Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

What does this allow us to do?



Resources are being wasted by not doing this



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

CrowdFlower (2016) Data Science Report