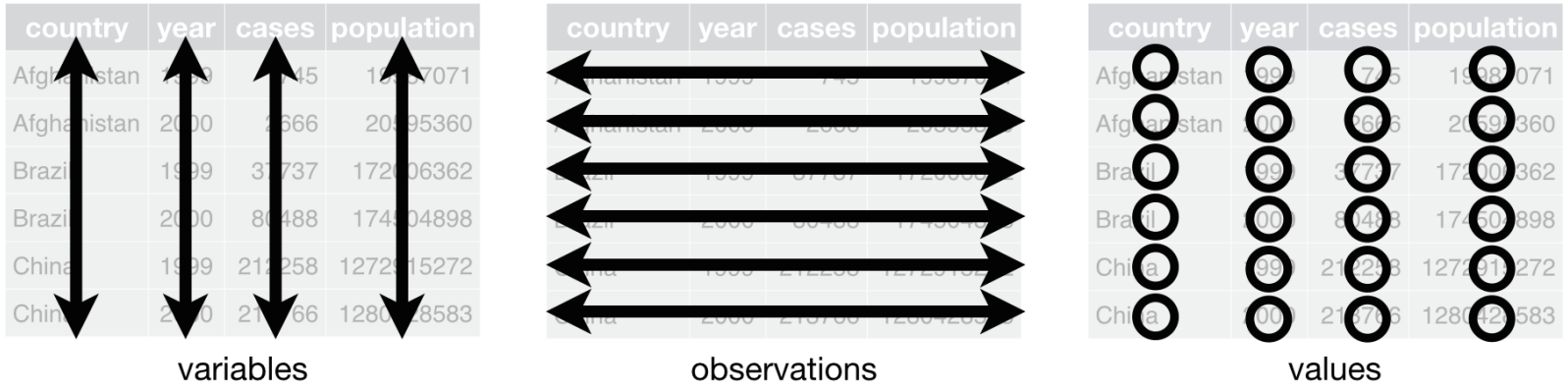
Tidy data

Francisco Rodriguez-Sanchez

https://frodriguezsanchez.net

Tidy data



https://r4ds.hadley.nz/data-tidy.html

Be careful with data entry and management in Excel!

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-

A. Hallmarks of well managed tabular data

1 Computer friendly

2 Descriptive headers

3 Atomized

4 Quality controlled

9 Data dictionary

sample_id	loc	habitat	temp	date	species	length_mm
13216	A	freshwater	15	2024-05-13	*Hypsibius dujardini*	0.3
98173	B	lichen	10	2024-06-01	*Milnesium tardigradum*	0.5
50232	C	soil	12	2024-05-06	*Echiniscus testudo*	0.4
36029	C	freshwater	18	2023-04-12	*Macrobiotus hufelandi*	0.6
61974	B	moss	14	2023-04-13	*Ramazzottius oberhaeuseri*	0.3
40079	A	lichen	11	2024-04-04	*Echiniscus testudo*	0.3
93823	A	soil	16	2024-05-17	*Milnesium tardigradum*	0.5
44467	C	freshwater	19	2024-05-16	*Hypsibius dujardini*	0.4
22896	B	moss	ND	2024-05-20	*Macrobiotus hufelandi*	0.6
83307	A	lichen	17	2024-05-17	*Ramazzottius oberhaeuseri*	0.3

sample_id: unique identifier for each sample
loc: collection site
habitat: collection habitat
temp: air temperature during collection (Celsius)
date: collection date
species: scientific name of specimen
length_mm: specimen length in millimeters

.csv
.tsv

10 Non-proprietary format

5 Defined null value

6 Date consistent

7 Read only copy

8 Analysis saved in separate file

B. Hallmarks of poorly managed tabular data

1 Colors as data

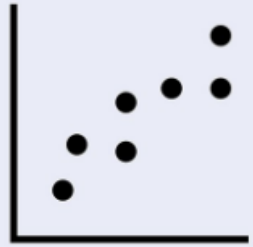
2 Headers not machine readable

3 Multiple data points per cell

4 Unvalidated data

9 Metadata in column header

Sample ID	Habitat and (Location)	°C	date	species	Length (mm)
13216	Freshwater (A)	15	05-13-2024	*Hypsibius dujardini*	0.31
98173	Lichen (B)	10	June 1 2024	*Milnesium tardigradum*	0.5
50232	Soil (C)	12	2024-05	*Echiniscus testudo*	0.4
36029	Freshwater (C)	18	2023-04-12	*Macrobiotus hufelandi*	0.6
61974	Moss (B)	14	2023-04-13	*R. oberhaeuseri*	300
40079	Lichen (A)	11	2024-04-04	*Echiniscus ??*	0.3
93823	Soil (A)	16	2024-05-17	*Milnesium tardigradum*	0.52
44467	Freshwater (C)	19	16-05-2024	*Hypsibius ??*	0.4
22896	Moss (B)		2024-05-20	*Macrobiotus hufelandi*	0.6
83307	Lichen (A)	17	June 17	*Ramazzottius oberhaeuseri*	0.3



.xls

10 Proprietary format

5 Undefined null value

6 Date inconsistent

7 Edited raw data

8 Analysis in the same file

Hertz & McNeill 2024

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- `http://www.datacarpentry.org/spreadsheet-ecology-lesson/`

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use Data validation in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- `http://www.datacarpentry.org/spreadsheet-ecology-lesson/`
- `http://kbroman.org/dataorg/`
- Broman & Woo: Data organization in spreadsheets

Common spreadsheet errors

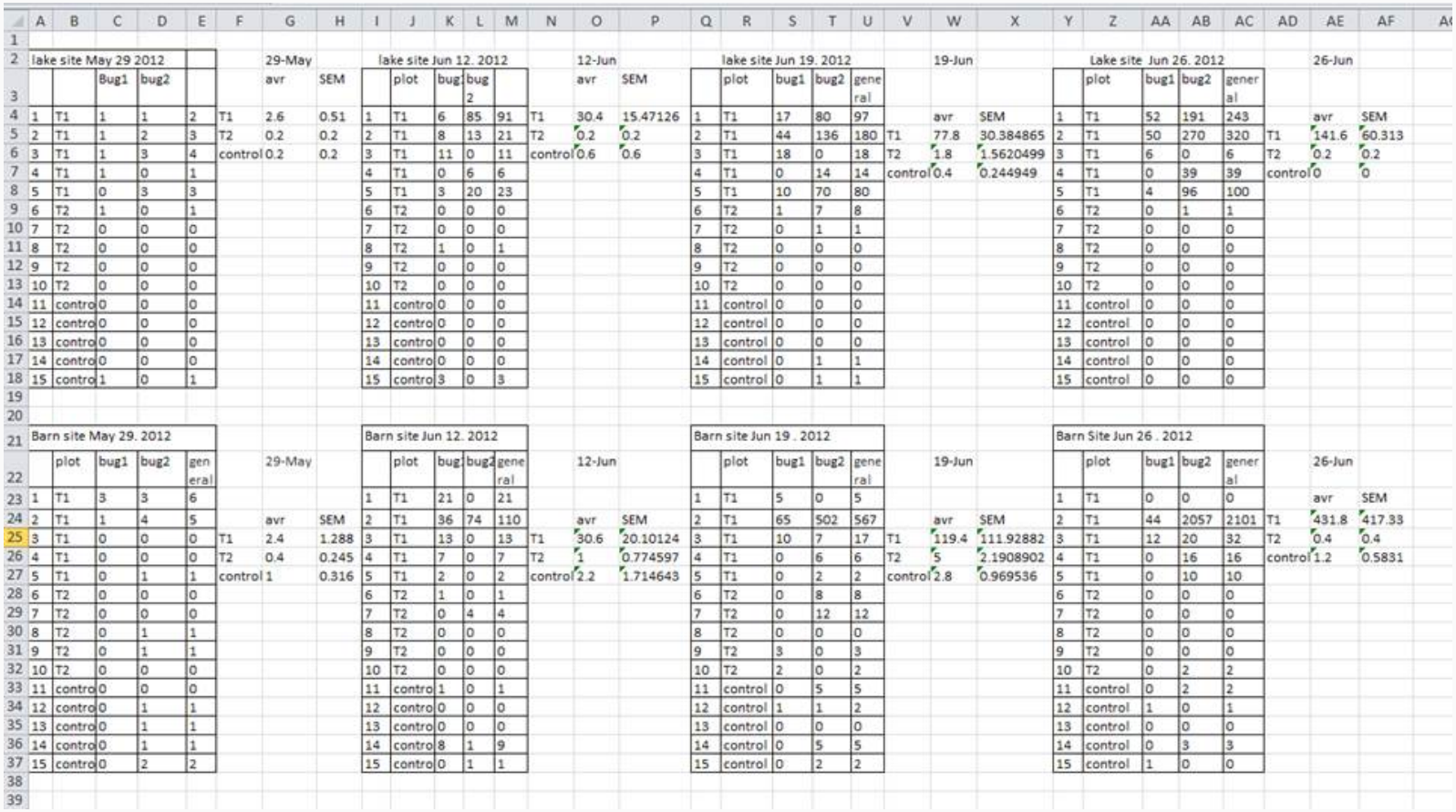
More than one variable per column

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Source: Data Carpentry

Multiple tables



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1																																
2	lake site May 29 2012						29-May		lake site Jun 12. 2012						12-Jun		lake site Jun 19. 2012						19-Jun		Lake site Jun 26. 2012						26-Jun	
3			Bug1	bug2			avr	SEM		plot	bug1	bug2			avr	SEM		plot	bug1	bug2	general					plot	bug1	bug2	general			
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	1	0	1				4	T1	0	6	6				4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	0	3	3				5	T1	3	20	23				5	T1	10	70	80				5	T1	4	96	100			
9	6	T2	1	0	1				6	T2	0	0	0				6	T2	1	7	8				6	T2	0	1	1			
10	7	T2	0	0	0				7	T2	0	0	0				7	T2	0	1	1				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	1	0	1				8	T2	0	0	0				8	T2	0	0	0			
12	9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0				10	T2	0	0	0			
14	11	contro	0	0	0				11	contro	0	0	0				11	control	0	0	0				11	control	0	0	0			
15	12	contro	0	0	0				12	contro	0	0	0				12	control	0	0	0				12	control	0	0	0			
16	13	contro	0	0	0				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
17	14	contro	0	0	0				14	contro	0	0	0				14	control	0	1	1				14	control	0	0	0			
18	15	contro	1	0	1				15	contro	3	0	3				15	control	0	1	1				15	control	0	0	0			
19																																
20																																
21	Barn site May 29. 2012						29-May		Barn site Jun 12. 2012						12-Jun		Barn site Jun 19 . 2012						19-Jun		Barn Site Jun 26 . 2012						26-Jun	
22		plot	bug1	bug2	general					plot	bug1	bug2	general					plot	bug1	bug2	general					plot	bug1	bug2	general			
23	1	T1	3	3	6				1	T1	21	0	21				1	T1	5	0	5				1	T1	0	0	0		avr	SEM
24	2	T1	1	4	5		avr	SEM	2	T1	36	74	110		avr	SEM	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10			
28	6	T2	0	0	0				6	T2	1	0	1				6	T2	0	8	8				6	T2	0	0	0			
29	7	T2	0	0	0				7	T2	0	4	4				7	T2	0	12	12				7	T2	0	0	0			
30	8	T2	0	1	1				8	T2	0	0	0				8	T2	0	0	0				8	T2	0	0	0			
31	9	T2	0	1	1				9	T2	0	0	0				9	T2	3	0	3				9	T2	0	0	0			
32	10	T2	0	0	0				10	T2	0	0	0				10	T2	2	0	2				10	T2	0	2	2			
33	11	contro	0	0	0				11	contro	1	0	1				11	control	0	5	5				11	control	0	2	2			
34	12	contro	0	1	1				12	contro	0	0	0				12	control	1	1	2				12	control	1	0	1			
35	13	contro	0	1	1				13	contro	0	0	0				13	control	0	0	0				13	control	0	0	0			
36	14	contro	0	1	1				14	contro	8	1	9				14	control	0	5	5				14	control	0	3	3			
37	15	contro	0	2	2				15	contro	0	1	1				15	control	0	2	2				15	control	1	0	0			
38																																
39																																

Could you avoid new tab by adding a column to original spreadsheet?

Using formatting, comments, etc to convey information

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

measurement device not calibrated

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Your turn: tidy up this messy dataset

https://ndownloader.figshare.com/files/2252083