

Introduction to data organisation

Stephen J Eglén
<https://sje30.github.io>
sje30@cam.ac.uk

Cambridge Computational Biology Institute
University of Cambridge
@StephenEglén

Slides: <http://bit.ly/eglen2019-data> (HTML) (CC-BY license)

Acknowledgements / Declarations

Data parasites everywhere; Senior Editor at *Scientific Data*.

(Last update Thu Jul 4 23:40:10 2019)

Introduction to reproducibility and reusability

Start with an interactive discussion. What data do you deal with; are you good at looking after it? what do you use?

What is **metadata**?

The reproducibility crisis

Many key findings in publications are either not independently verified, or fail verification when it is attempted (Baker, 2016).

Duke oncogenomics scandal. Awesome detective work by Keith Baggerley and Kevin Coombes. <https://www.youtube.com/watch?v=7gYIs7uYbMo>

Disclaimer: do I mean "reproducibility" or "replicability"? (Barba 2018)
<https://arxiv.org/pdf/1802.03311.pdf>

Moral or selfish approach?

Markowetz *Genome Biology* (2015) 16:274
DOI 10.1186/s13059-015-0850-7



COMMENT

Open Access

Five selfish reasons to work reproducibly



Florian Markowetz

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

Keywords: Reproducibility, Scientific career

how science actually is. And, whether you like it or not, science is all about more publications, more impact factor, more money and more career. More, more, more... so how does working reproducibly help me achieve more as a scientist.

Reproducibility: what's in it for me?

In this article, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

Selfish reasons to share

Why not align what is good for science with what is good for scientists?

1. Funding mandates (REF + enforcement from Wellcome Trust)
2. Credit through data papers
3. Fixes data bugs / errors in analysis
4. Prevent data loss ([Vines et al 2014](#)). e.g. students have a habit of leaving...
5. Your future self is probably one of the main beneficiaries of sharing.
6. *Now* is a very good time to be an open scientist.
7. Leads to further collaborations
8. Reviewers can do more work...

Rule 1: Data should be shared

Given the cost of generating data, I think data relating to a publication should be shared along with the paper.

Funders (and increasingly many, but not all, journals) agree.

The Lancet

Longo DL, Drazen JM (2016) Data Sharing. N Engl J Med 374:276–277
<http://dx.doi.org/10.1056/NEJMe1516564>.

Data sharing sounds great ... but

1. Someone not involved in collecting data may misunderstand it.
2. "a new class of research person will emerge --- people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as **research parasites**"."

Response from the community was to embrace the term "research parasite", e.g. <https://twitter.com/dataparasite>

This is just good science. I think there are few valid reasons for not sharing data.

FAIR data standards

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

<https://www.nature.com/articles/sdata201618>

Writing a good data management plan (DMP)

Michener WK (2015) Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol 11:e1004525
<http://dx.doi.org/10.1371/journal.pcbi.1004525>.

Things to be wary of

GDPR, Ethics, anonymisation.

Getting help

- CambridgeData Champions (phd students, postdocs, faculty)
<https://www.data.cam.ac.uk/data-champions-search>
- Office for Scholarly Communication <https://osc.cam.ac.uk/>
- UK Reproducibility Network (UKRN)
<https://www.bristol.ac.uk/psychology/research/ukrn/>
- The Turing Way <https://github.com/alan-turing-institute/the-turing-way>

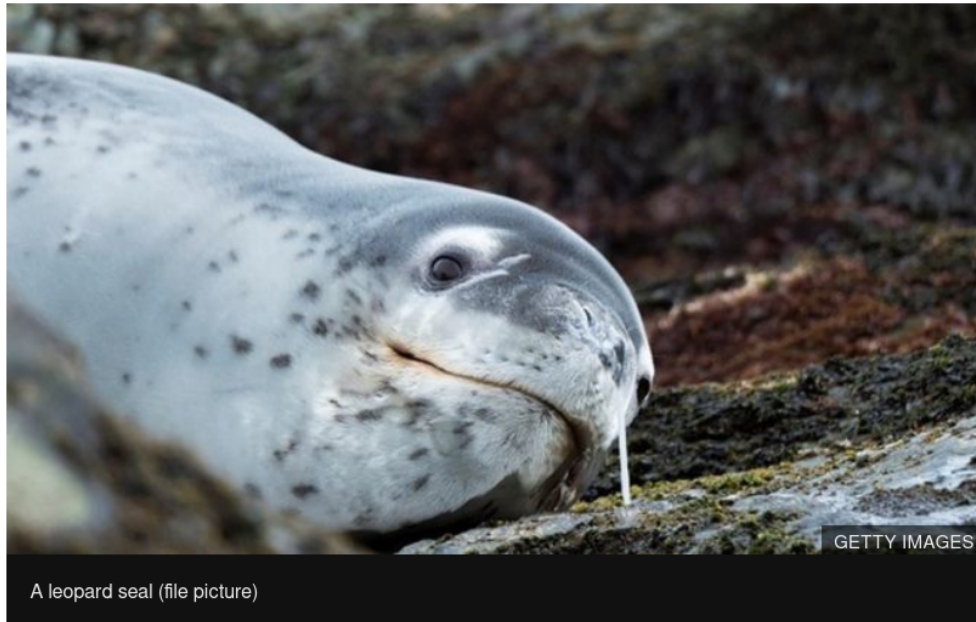
Losing your data

We all know the stories of PhD students losing their laptop (or in my case, a server), and along with it their only copy of their data. Sometimes data comes back though:

Memory stick found in frozen seal faeces in New Zealand

🕒 5 February 2019

[f](#) [💬](#) [🐦](#) [✉](#) [Share](#)



A slab of seal poo used for scientific research in New Zealand has led to the unlikely discovery of a USB stick full of holiday snaps.

Licence

What restrictions do you want to put on your data? If there is no licence, people cannot easily reuse data.

CC0 https://wiki.creativecommons.org/wiki/CC0_use_for_data

Don't edit the raw data files.

Keep pristine (e.g. read-only) in own folder. Write scripts to manipulate your data.

Open formats

- Data and metadata best stored in open formats
- What does **open** mean?
- Adopt conventions from your field if there any.
- Else use things like HDF5 (matlab), CSV, XML, SQL, JSON (...)
- Spreadsheets better as .csv (not .xls or .xlsx) if used.

Spot the issues

S1Sh.cuo					
	A	B	C	D	E
1		Group1	Group2		
2		Day 0			
3	Sodium	139	142		
4	Potassium	3.3	4.8		
5	Chloride	100	108		
6	BUN	18	18		
7	Creatine	1.2	1.2		
8	Uric acid	5.5*	6.2*		
9		Day 7			
10	Sodium	140	146		
11	Potassium	3.4	5.1		
12	Chloride	97	108		

A better version

Table_S1_Shanghai_blood.xls						
	A	B	C	D	E	F
1	Parameter	Day	Control	Treated	Units	P
2	Sodium	0	139	142	mEq/l	0.82
3	Sodium	7	140	146	mEq/l	0.70
4	Sodium	14	140	158	mEq/l	0.03
5	Sodium	21	143	160	mEq/l	0.02
6	Potassium	0	3.3	4.8	mEq/l	0.06
7	Potassium	7	3.4	5.1	mEq/l	0.07
8	Potassium	14	3.7	4.7	mEq/l	0.10
9	Potassium	21	3.1	3.6	mEq/l	0.52
10	Chloride	0	100	108	mEq/l	0.56
11	Chloride	7	97	108	mEq/l	0.68
12	Chloride	14	101	106	mEq/l	0.79

Dates


PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. $2\frac{7}{2}$ -13 2013.158904109
MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ 2013
10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ & 5 & 6 & 7 & 8 \end{matrix}$ 

Source: <https://xkcd.com/1179/>

Spreadsheet concerns

Ziemann M, Eren Y, El-Osta A (2016) Gene name errors are widespread in the scientific literature. *Genome Biol* 17:177 <http://dx.doi.org/10.1186/s13059-016-1044-7>.

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

For example, gene symbols such as SEPT2 (Septin 2) and MARCH1 [Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase] are converted by default to '2-Sep' and '1-Mar', respectively.



File names

```
paper_final_version.docx  
paper_final_version_edits.docx  
paper_final_version_edits2.docx  
paper_final_version_really_final.docx
```

Opinions vary ... Little bit of metadata good:

```
jan15_ctl.txt  
jan15_drug.txt
```

But how much metadata can you cram into

Some mechanics: avoid spaces, special characters (space, parens, punctuation)

Learn about regular expressions

Further spreadsheet advice

Broman KW, Woo KH (2018) Data Organization in Spreadsheets. Am Stat 72:2–10 Available at: <https://doi.org/10.1080/00031305.2017.1375989>

- Be consistent. Write for robots, not humans.
- Choose good names for your variables (column headings)
- Make spreadsheet a dense rectangle (no holes); *tidy* data.
- No calculations in raw files
- Use data validation to avoid errors
- Save copies in plain text

Ten simple rules for the Care and Feeding of Scientific data

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y, Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A (2014) Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol 10:e1003542
<http://dx.doi.org/10.1371/journal.pcbi.1003542>

1. Love your data, and help others love it, too
2. Share online with permanent identifier
3. Conduct science with a particular level of reuse in mind
4. Publish workflow as a context
5. Link your data to publications as often as possible
6. Publish your code (whatever state is it in)
7. State how you want to get credit
8. Foster/use data repositories
9. Reward colleagues who share data
10. Be a booster for data science

Find a decent repo

Guidelines from PLOS and Scientific Data

<https://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories>

<https://www.springernature.com/gp/authors/research-data-policy/repositories/12327124>

General repo: Zenodo / figshare

Many universities have own institutional repository, but I find this the wrong way to cut the cake (prefer by subject, not location)

Next steps

- Data verification
- Provenance
- Data mining (the robots *are* coming)