

Open and Reproducible research: the new frontier

Dasapta Erwin Irawan, R. Willem Vervoort & Gene Melzack

19 January 2018

Open data and reproducible research

Summary

- ▶ Data and **open** data
- ▶ The issue with verifiable research
- ▶ Open and reproducible research
- ▶ Challenges

Data and **open** data (1)

How many of you have?

- ▶ Data from old research on a floppy disk, zip disk, usb stick, mobile harddrive: *I am going to publish that one day!*.
- ▶ Data on your harddrive from your PhD student, but you have no idea how she/he organised it
- ▶ received data from a colleague and spend hours reformatting it to your needs
- ▶ asked data from a colleague, who said yes, but then could not find the data
- ▶ spend hours filling in forms and e-mailing with another institution to access data (climatology!!)

Data and **open** data (2)

Have you ever experienced any of the following?

I have got this great idea,

- ▶ but I cannot access the data
- ▶ but I can't find a simple example of how to do the analysis correctly
- ▶ but my model won't run without this specific data

or... I read this great paper,

- ▶ but I think the analysis is wrong.
- ▶ but I can't work out how exactly they did this analysis
- ▶ but I think I know how to take the next step if I could use the data

Data and **open** data (3)

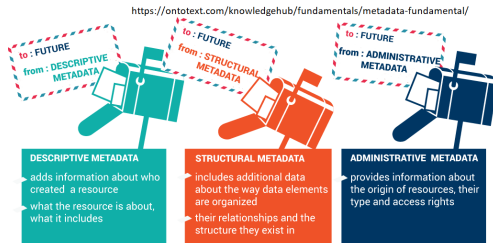
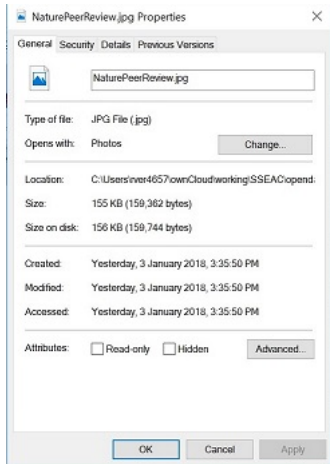
or... I am using this piece of software,

- ▶ but I don't understand how the algorithm exactly works
- ▶ but I would like to change it slightly to work better for my research
- ▶ but I can't access the code without a hefty fee

Data and **open** data (4)

- ▶ Data can be anything, it can be words, numbers, pictures, even bits of code or algorithms
- ▶ Most data is currently difficult to access
 - ▶ individuals computers
 - ▶ protected
 - ▶ not well described
- ▶ **Open data** is not only easily accessible, but is also well described
 - ▶ it has all the meta data to describe the *provenance* and the *characteristics*
- ▶ Examples are data from the IPCC and NOAA
- ▶ We will look at this in more detail later

Meta data



Vocabulary and data dictionary

- ▶ The keywords for your metadata should originate from a **vocabulary** or an **ontology**
 - ▶ Ontology: set of controlled terms for keywords with a hierarchy: example [FOR codes](#)
 - ▶ A good ontology would have related terms, Wikipedia is an example of a system that uses an ontology
 - ▶ A vocabulary is a more simple list of keywords, for example, most journals require you to choose from specific keywords when you submit a paper

Vocabulary and data dictionary

- ▶ Data dictionary is simpler:
 - ▶ describes columns in a data sheet
 - ▶ describes layout of code structure
 - ▶ describes files and folders
- ▶ example data dictionary: [Readme file in example project](#)

The issue with verifiable research

- ▶ The current process is peer review
- ▶ Requires knowledgeable reviewers
- ▶ System has been questioned recently, can it be fair and can it be maintained?
 - ▶ Fairness to different languages (non English)
 - ▶ Cost of traditional publishing
 - ▶ Hidden cost of reviewer's labour
- ▶ open and reproducible research might be a solution



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 532 > Issue 7599 > Comment > Article

NATURE | COMMENT

Peer review: Troubled from the start

Alex Csizsar

19 April 2016

Pivotal moments in the history of academic refereeing have occurred at times when the public status of science was being renegotiated, explains Alex Csizsar.

[PDF](#) [Rights & Permissions](#)

Subject terms: [Peer review](#) · [Publishing](#) · [History](#)

Referees are overworked. The problem of bias is



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 530 > Issue 7588 > Comment > Article

NATURE | COMMENT

Reproducibility: A tragedy of errors

David B. Allison, Andrew W. Brown, Brandon J. George & Kathryn A. Kaiser

03 February 2016

Mistakes in peer-reviewed papers are easy to find but hard to fix, report David B. Allison and colleagues.

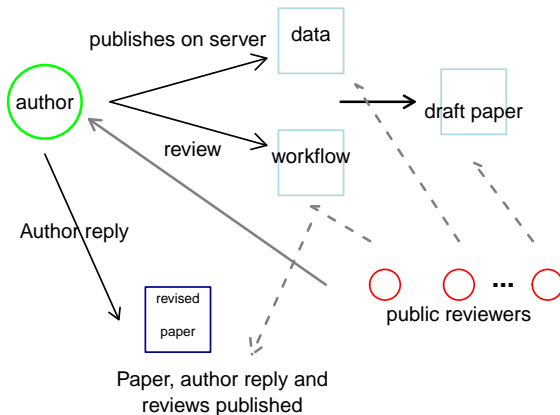
[PDF](#) [Rights & Permissions](#)

Subject terms: [Communication](#) · [Publishing](#) · [Peer review](#)

An ideal description of open, reproducible, peer reviewed research

What would ideal open, reproducible research look like?

- ▶ all data and analyses should be open and accessible



The roadblocks to open data and reproducible research

Why is this not happening?

- ▶ Skill and ability to publish open data and workflows (researcher)
 - ▶ meta data
 - ▶ workflow documentation
- ▶ Provision of infrastructure (institution)
- ▶ IP and ownership claiming
- ▶ unbiased reviews and internet trolls



Three major components to reproducible data and research

- ▶ Open and accessible data
- ▶ For *raw* data: fully documented metadata (what the data actually is, and how it was generated or measured)
- ▶ For *derived* data: fully described and documented workflow (*provenance*, how the data was manipulated)



New skills that we need to make it happen

- ▶ How do we regularly and consistently describe metadata with our data
- ▶ How do we easily publish data and preprints (How does our institution manage this)
- ▶ Understanding how we can protect IP: licencing and digital identifiers
- ▶ Getting recognition and support from our institutions for open data publications



How this workshop fits in

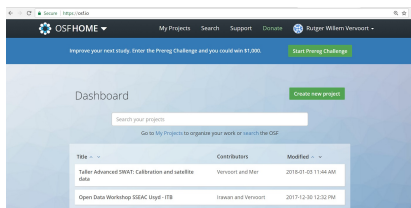
Over the next 4 days, we will teach you about:

- ▶ data, and how to write good metadata
- ▶ netcdf and why this might be useful
- ▶ workflows and how to record a workflow using Rmarkdown
- ▶ code and how to manage code via github
- ▶ how to get recognition, DOI and licences

What is already out there?

There is already a lot out there! (although not everything is free)

- ▶ Data journals, for example Data in Brief and Data, and there is a growing list
- ▶ Data repositories, for example PANGAEA, but here is a long list
 - ▶ The University of Sydney also runs its own data repository
- ▶ Journals to publish workflows, for example MethodsX
- ▶ Full open science repositories, such as Zenodo and OSF



Class activity (15 minutes)

- ▶ Discuss in groups:
 - ▶ How you have shared data in the past?
 - ▶ What are the main ways how you currently store and *curate* data. How easy would it be for someone else to access your data?
 - ▶ What actions do you have to take to share data.
 - ▶ How easily have you accessed someone else's data?
- ▶ As a group report back to summarise