Open Science: principles and skills

Dr. Julien Colomb

About this document

This document tries to gather information about how to become a better (open) scientist and follow the highest standard in data management and dissemination. The audience is primarily PhD students, but more experienced researchers may also benefit from it. Importantly, we present what should be done independently from how computer tools may help. In this respect, the document is meant to be read as a **pdf**, while it is written with **rmarkdown**.

I am looking for help: co-authors, reviewers and commentators are welcome, see https://github.com/jcolomb/opensciencesskills.git to see how you could help.

Introduction

New standards in science

Standards are continuously changing. In 2007, I (Julien Colomb) published a paper about the anatomy of the Drosophila gustatory system having one good preparation per genotype. Today, I would need 10 of these. At the time, we (the community) thought differences between specimens were due to technical reasons, such that the best preparation was giving the "real" result. Nowadays, we know that the variability is biological and not technical, and one needs more preparation to estimate the "average result". We know more about the system, which asks more from us in return. **Standards are changing and it is good.** But it also means more work, leading to a need for more collaboration. Similarly , open science is a new standard which is meant for achieving better science.

With more pressure on scientists, the reproducibility of science is becoming problematic. Trust in science is therefore fading. To regain that trust, scientists need to change their workflow: perform better data analysis and **publish both their data and their analysis (open data)** together with their findings. Accordingly, early information exchange becomes a standard, since it leads to faster and more reliable scientific discovery. Scientists are encouraged to **share their ideas and results prior to publication (open science)**. Although open scientists are still a minority, the new generation of scientists should prepare themselves for this new paradigm.

In order to achieve this, there is a need for:

1. An adapted scientific workflow
2. Novel (computer) skills to run it efficiently

In January 2016, I had a rapid discussion on twitter: a scientist was asking about what new skills she should teach to her PhD students (who were writing their thesis). I told her about using latex, version control and R. She was unaware of any of these subjects and it made me wonder. I decided to write this letter. Each section will be dealing with (1) basic solutions and (2) software and computer skills that can make things easier and faster.

In practice: open means better managed

You may have heard of the cost of being open, and discussion about who should pay for these additional costs. In this document, we will also try to demonstrate that we should not talk about costs, but about

investments. As a singular scientists, it means investing time in aquiring computer skills, and investing time in data and information management prior to data gathering.

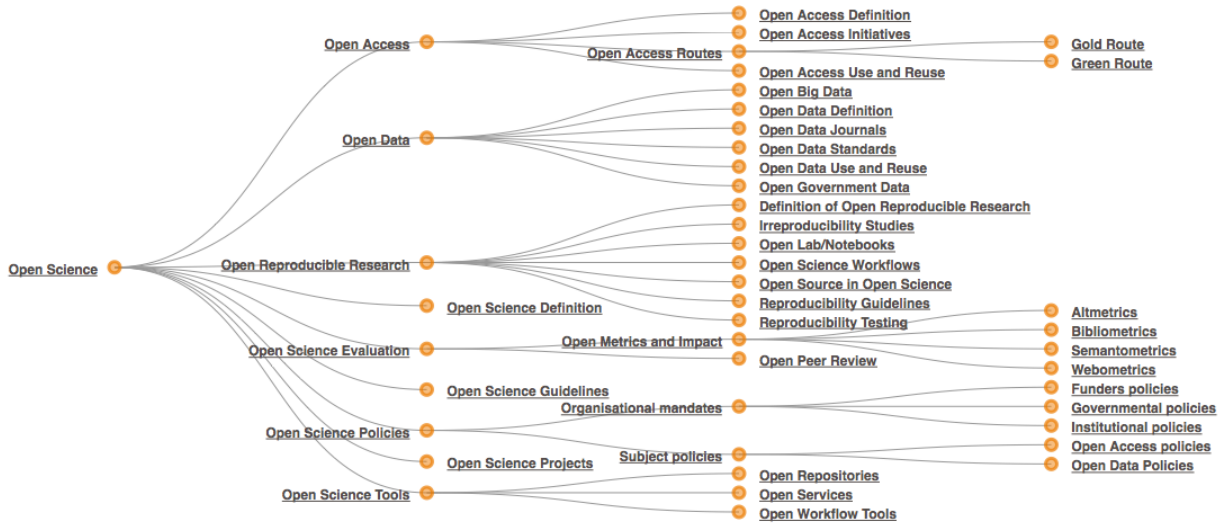
> I did my PhD in 2007. At the time, no one was speaking about open data or open science, but my boss was telling me: Your notebook should be readable in 5 to 10 years from now. I was quite bad at following the advice and everything I did before embracing the open data movement is probably not viable anymore, even if I could find my notebook again. What has changed with open data? Time proximity: while putting your data in the open (or preparing it to be so), you need to make it understandable by all (this includes yourself in 10 years). I am just not allowed to be lazy anymore and I have to manage my data. Interestingly, when done correctly, it does save time: the (time) cost of managing your data is counterbalanced by the effectiveness of your work: the data becomes easier to find, easier to analyse and easier to re-analyse. The 10 hours used to get my data correctly labeled saved me 50 hours of searches.

Open Science

> Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional. (https://en.wikipedia.org/w/index.php?title=Open_science&oldid=699123667)

It comprises different areas, i.e. access to the scientific literature and data upon publication, open evaluation of scientific policies (see below).

Fig.1 : from fosteropenscience, it shows the different areas of interest where open science is at play



topics (working toc)

most inspiration coming from https://github.com/kbroman/steps2rr

1. Data gathering
    - sampling, randomisation and blindness

- tidy table
- Use of spreadsheets

2. Data management
    - data organisation: raw data files as bases
    - use a master_file
    - filenames
    - publish data
    - use the same rules for code
3. Data analysis
    - Documentation of the analysis
    - The fishing problem, the registration solution.
    - Using R
4. Version control
    - filenames and versioning
    - documentation
    - Git
5. Publication
    - Where, when, how
    - licences
    - Automation

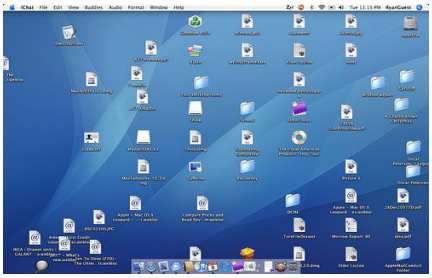
Open data has no cost

Organizing data for it to be shared represents a lot of work. This is often presented as an extra cost for scientists, there would be a cost to share own's data. I do not agree: there is no cost in making own's data sharable, but there is a time **investment**. An analogy will best show my point: The very basic rule in a molecular biology lab is: tidy your bench, make sure that your solutions, samples and equipment are kept in the right place and temperature and that the equipment is well maintained. It will save you time for your next experiment (finding the right reagents and your tools). The same apply for data: **tidy your data files**, make sure that your data is kept in the right place and format and that your analysis tools are well maintained.

The amount of time saved by organizing your data overwhelm by far the time invested to organize it. Everybody aggrees that the extra cost to share organized data is tiny. In conclusion, Open data is not costly, it is only a question of training and getting the right habits.

You won't let your bench messy.

Why do you let your files be?



Tidy up you data and analysis code!

That's just efficient scientific practice