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 could publish the anatomy of the Drosophila gustatory system having one good preparation per genotype. Today, I would need 10 of these. At the time, we thought differences between specimen 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 need more preparation to estimate the average result. We know more about the system, which ask 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. Open science is the mean to achieve better science.

With more pressure on scientists, the reproducibility of science is becoming problematic. Trust for science is therefore fading. To regain that trust, scientists needs to change their workflow: make 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 idea and results prior to publication (open science)**. Although incentive mechanisms are not yet in place, the new generation of scientists should make them ready for this new paradigm.

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

1. An adapted scientific workflows
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 bring to her PhD student (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

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

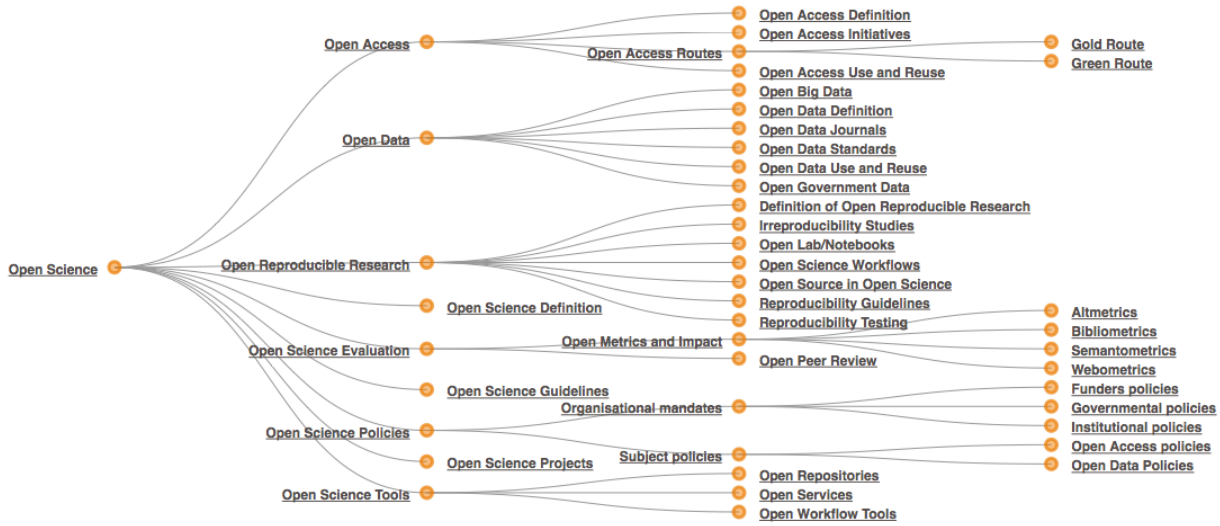
the advice and everything I did before embrassing the open data movement is probably not viable anymore, even if I could find my notebook back. What has changed with open data? Time proximity: while putting your data in the open (or preparing it to be so), you need to make it understandable by all (this 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 save 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 to scientific policies (see below).

Fig.1 : from fosteropenscience, it shows the different areas of interest where openscience 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