

Open Science In Practice



Lessons learnt and application

Etienne Thalmann, October 2017

Based on presentations during EPFL Open Science in Practice 2017 summer school by:

Benedikt Fecher

Arnaud Vaganay

Martin Vetterli

Laurent Gatto

Jessica Polka

Lawrence Rajendran

Kirstie Whitaker

Sunje Dallmeier-Tiessen

Marta Teperek

Lucia Prieto

Victoria Stodden

Gaël Varoquaux

Tim Head

Michel Jaccard

Dasaraden Mauree

All slides are available under CC-BY licence, unless indicated otherwise on specific slides



Why do we need to open Science?



Open science is not a new concept. Scientists have been sharing their knowledge for centuries.

But today much of the science is inaccessible or published in a way that lacks transparency.

- ▷ 4-5 publishers own most of the journals and hide the articles behind paywalls. This system prevails (inertia)
- ▷ The goals have changed. It's mostly about citation metrics. *“When a measure becomes a target, it ceases to be a good measure”* (Goodhart's Law)
- ▷ Theoretical work or negative results are hard to publish. It needs to be a story!
- ▷ A lot of the published research is never reproduced because the information necessary to do so is missing

The opinion of EPFL president Martin Vetterli

For universities, open science is important because:

- ▷ It increases visibility
- ▷ The research is paid by taxpayers (public money) and it should thus be accessible to them
- ▷ Open science increases the quality and efficiency of publications
- ▷ The impact of research is larger (more people have access to it)

For a scientist the benefits include

- ▷ Increased citation due to openness
- ▷ Better science due to reproducibility and early feedback
- ▷ Easiness to build on his own work
- ▷ An open showcase of knowledge and skills for grant applications, future employers, potential collaborators etc.

Some common barriers to opening science

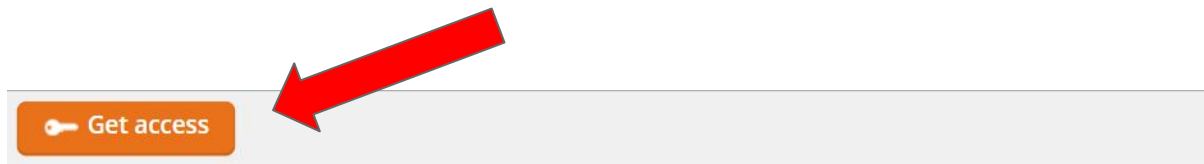
- ▷ Researchers are afraid of being scooped
- ▷ They don't want to share low quality results, are afraid of criticism
- ▷ They don't know what to share or how to share it
- ▷ Ethical reasons (ex: personal data from patients)
- ▷ Intellectual property

Publishing

The classical publishing format

The cost for one publisher is about 5 million dollar per university per year

Knowledge is hidden from people who cannot afford it.
Ex: 1858 Darwin paper behind a paywall



[Explore this journal >](#)

On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection.

Charles Darwin Esq., F.R.S., F.L.S., &c F.G.S., Alfred Wallace Esq.

First published: August 1858 [Full publication history](#)

DOI: 10.1111/j.1096-3642.1858.tb02500.x [View/save citation](#)

The publishing contract

Do you own the rights on the published article?

In some cases no, you give the full authorship rights to the publisher. You have to ask them if you want to re-use it

Which version can you reuse?

- ▷ **Pre-print** (before submission to the journal)
Should always be the case, the publisher didn't do any work and it belongs to you
- ▷ **Post-print** (after reviewing, without publisher formatting)
Sometimes yes, for in-house teaching purpose
- ▷ **Published version**
You cannot re-use it at all with some journals (ex: elsevier)

Open access models

Gold open access

- ▷ The author pays so that his paper is published in a journal that is freely available for any reader

Hybrid model




- ▷ The author pays so that his paper is published in open access in a journal where only part of the articles are in open access
- ▷ Not recommended by the EPFL library. EPFL pays to publish and pays again anyway the subscription to the journal to have access to the other articles




Green open access

- ▷ The preprint and postprint are openly readable. An embargo (can go from 0 to 3 years) or reuse conditions can apply

Example: Journal of Mechanical Design vs. Science

Comparisons using SHERPA/RoMEO

Journal:	Journal of Mechanical Design (ISSN: 1050-0472, ESSN: 1528-9001)
RoMEO:	This is a <u>RoMEO white</u> journal
Paid OA:	A paid open access option is available for this journal.
Author's Pre-print:	 author cannot archive pre-print (ie pre-refereeing)
Author's Post-print:	 author cannot archive post-print (ie final draft post-refereeing)
Publisher's Version/PDF:	 author cannot archive publisher's version/PDF

Journal:	Science (ISSN: 0036-8075, ESSN: 1095-9203)
RoMEO:	This is a <u>RoMEO green</u> journal
Author's Pre-print:	 author can archive pre-print (ie pre-refereeing)
Author's Post-print:	 author can archive post-print (ie final draft post-refereeing)
Publisher's Version/PDF:	 author cannot archive publisher's version/PDF

Open journals

Pros

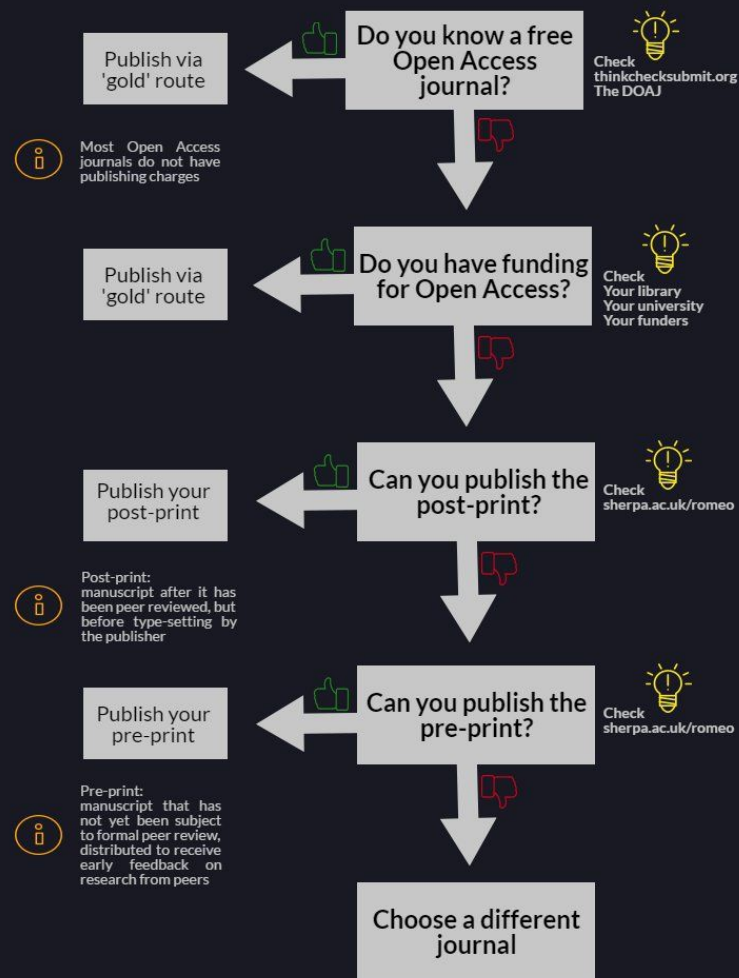
- ▷ More citations
- ▷ A further dissemination of knowledge
- ▷ They force you to publish the data (reproducibility)

Cons

- ▷ The price
- ▷ They are not recognized in the conventional crediting systems
Solution: use alternative systems to evaluate the impact of a single article (ex: Altmetric, PLOS shows the amount of sharing/viewing/citing)

HOW TO MAKE YOUR RESEARCH OPEN ACCESS

FOR FREE AND LEGALLY



How to use your article in your thesis?

- ▷ Some publishers don't consider a thesis as a publication -> It's ok to reuse your article
- ▷ Publish in "gold model" -> pay beforehand so that the work has a cc-by license
- ▷ Publish beforehand under an open license the part of the work that you will want to reuse
- ▷ Put your work in a preprint
- ▷ Negotiate with the journal beforehand to have an agreement to allow reuse in your thesis

Preprints

Preprints

- ▷ What are they?

A draft (incomplete or final) of an article which has not yet been submitted for publication and peer review

- ▷ Qualities

- Open-access
- Permanent
- Versioned (has a timestamp)
- Citable

- ▷ Where to publish/find them ?

- OSF preprints: <https://osf.io/preprints>
- engrXiv Preprints: <https://engrxiv.org>
- ...

Preprints

Why publish them?

- ▷ To avoid hiding your research for a while before publication
- ▷ To prove at what time you had an idea in case of a delay in publication due to peer review
- ▷ To have the possibility to reference not yet published papers in your grant proposals, CV, articles etc.
- ▷ To have immediate feedback from the scientific community
- ▷ To be visible to some editors which might offer to publish the research
- ▷ To find potential co-authors

Can I still publish in a journal?

Many journals allow pre-prints. 5 years after posting, more than 70% of arXiv preprints appear in journals.

How to check if a journal allows preprints?

- ▷ Wikipedia -> List of academic journals by preprint policy
- ▷ SHERPA/RoMEO: Search engine for Publisher copyright policies

Licenses

Usage

The license only explicit what you can do without asking the author. You can always ask the author for an exception.

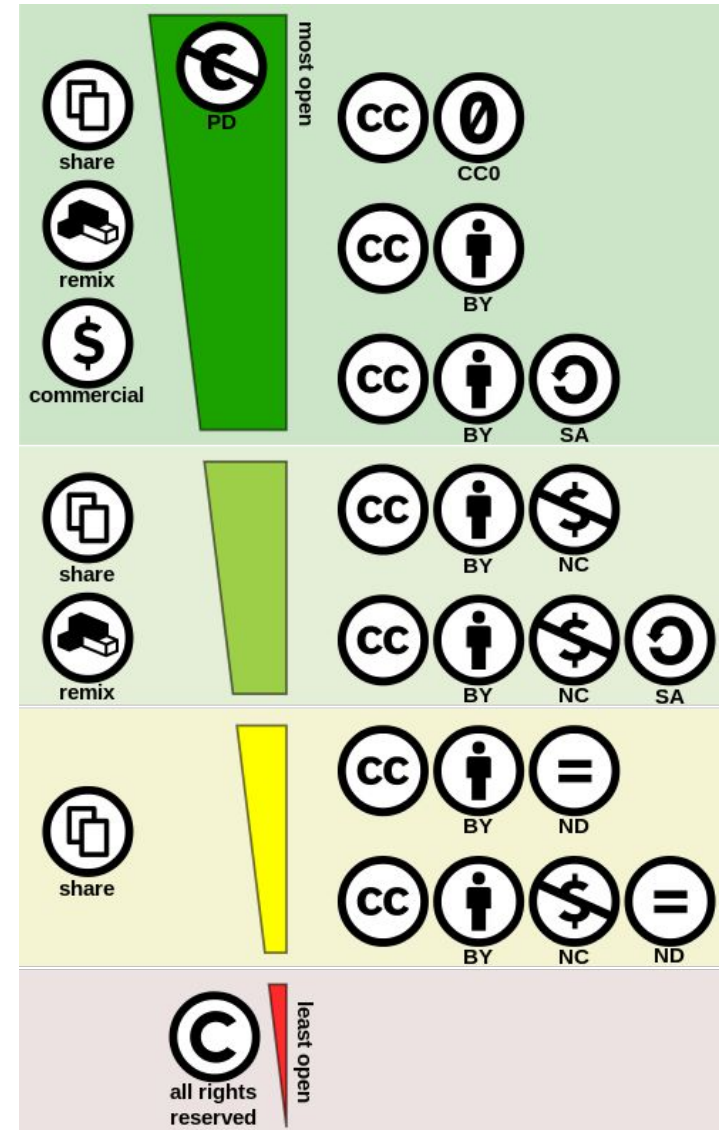
Ex: if there is an NC (non commercial) license, you can contact the author to negotiate a commercial use.

Without any specification, the standard copyright applies. You need to ask the author for anything you would like to do with his work.

Creative Commons license

CC

- BY -> need to credit the author
- SA -> need to put the same license on derived products
- NC -> non-commercial use
- ND -> no derivative works



GNU General Public License (GPL)

It gives the freedom to run, share and modify the software.

But it specifies that any derivative project has to stay open and have the same license (share-alike).



Licenses for code

CC licenses do not apply, they are for creative works. You can use the following licenses which do not have the share-alike limitation.

- ▷ **MIT license**

A permissive free software license with very few limitations on re-use and sharing. Compatible with other free software licenses.

- ▷ **BSD license**

Berkeley Software Distribution. A family of permissive free software licenses.

RRS (reproducible research standard)

Victoria Stodden, Stanford university

Instead of starting a new license, RRS is a suite of recommendations using the existing one.

They recommend using:

- ▷ CC BY for text, figures, etc
- ▷ Modified BSD, MIT license or similar for code
- ▷ CC0 for data (public domain)

This is because with any other license, if someone wants to re-use the data, mix it with other data and compute some result, they have to specify exactly which results come from which sets of data. This makes it very complicated to use the data. Scientific ethic dictates that you will be cited as author of the data

Reproducibility



Reproducibility

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

credit: Kirstie Whitaker

Advantages of reproducibility

1. It helps to avoid disaster

- Ex: Data analysis problems having lethal impacts have been spotted in a series of high-impact breast cancer research articles. At the time of the discovery, patients in clinical trials were treated on the basis of these results. The problems would have been spotted easily if the data had been available and the analysis transparent.

2. Reproducibility makes it easier to write papers

- If the data changes, all the results can be updated easily. You can have more confidence that results are up to date and more eyes can verify them easily.

Sources:

- Florian Markowetz. Five selfish reasons to work reproducibly. *Genome Biology* **16** Springer Nature, 2015
- Keith A. Baggerly, Kevin R. Coombes. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics* **3**, 1309–1334 Institute of Mathematical Statistics, 2009.

3. It helps reviewers see it your way

- It enables the reviewers to truly understand the research and propose valuable changes. They can even test their idea on the data before suggesting it to you.

4. Reproducibility enables continuity of your work

- Avoid situations where you do not remember what you did a few months ago or cannot find the data from a previous experiment, or have to continue working on the project of a colleague who left and have a hard time understanding what he did.

5. You will gain the reputation of being an honest and careful researcher

6. You will gain confidence when showing or defending your results

Making reproducible experiments and replicating it yourself with a validation set gives you a lot of confidence

7. You will make you save time in the long run

- It might take time to tidy up the data and analysis at first but you will save a lot of time when you will look at it later or continue working on it.

Tips to make one's research reproducible?

- ▷ Document well your data and code, write documentation as you go along
- ▷ Keep your project organized and easily accessible, name your files and directories in some informative way
- ▷ Store your data and code at a single backed-up location
- ▷ Reproducibility is improved when there is less clicking and pasting and more scripting and coding
- ▷ Add the information for replicating in a the "supplementary" of the paper. It does not need to make your paper heavy

Data sharing



Why?

- ▷ In order to publish in major journals, authors almost always have to deliver their data
- ▷ It is the first step to making your research reproducible. You still need to make the analysing method or code available
- ▷ In order to easily find your data some years later

Where?

On a repository. It can be found using the re3data.org repositories database

One interesting repository is [Zenodo.org](https://zenodo.org)

- ▷ Hosted by CERN
- ▷ Free data submission for any research as long as it is openly published
- ▷ Repository for data, code, slides and more
- ▷ Allows to have a DOI for your data
 - Citable in your paper
 - Always available in the state it was when you published
 - Citable by other papers
 - Easy to give to people who contact you to have access to your data

Python and Jupyter



An interesting open source alternative to MATLAB

Pros

- ▷ Free
- ▷ Can do almost everything you can do with MATLAB
Ex: Pandas library for data analysis, Matplotlib for plots similar to Matlab
- ▷ General-purpose language
Used for scientific computing, enterprise software, web design etc.
- ▷ Your code is portable
Someone who doesn't have a Matlab license can also run it. You can use it if you change employer
- ▷ Easy to read, beautiful programming language
(I don't have a personal opinion about this yet).
- ▷ Can easily find help from the community on the internet
- ▷ All the algorithms are transparent

Cons

- ▷ Need to learn a new programming language
- ▷ Does not have Simulink
- ▷ Need to install extra packages
- ▷ Not widely used at EPFL
- ▷ Does not have some of the powerful MATLAB toolboxes and solvers
- ▷ Does not have the friendly MATLAB environment

Sources:

http://www.pyzo.org/python_vs_matlab.html
<https://stevetjoa.com/305/>

Jupyter notebook

“The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.” – <http://jupyter.org/>

- ▷ Can mix text, code (ex: Python) and plots/animations
- ▷ Allows you to run and edit scripts on a webpage and display the output

Jupyter notebook - an example

Visualize the results

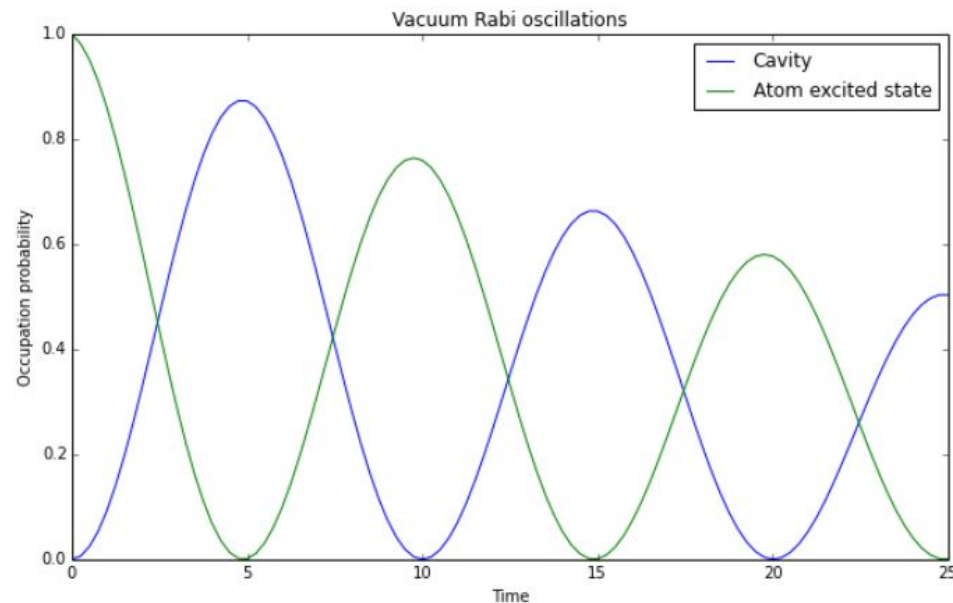
Here we plot the excitation probabilities of the cavity and the atom (these expectation values were calculated by the `mesolve` above). We can clearly see how energy is being coherently transferred back and forth between the cavity and the atom.

```
In [8]: n_c = output.expect[0]
n_a = output.expect[1]

fig, axes = plt.subplots(1, 1, figsize=(10,6))

axes.plot(tlist, n_c, label="Cavity")
axes.plot(tlist, n_a, label="Atom excited state")
axes.legend(loc=0)
axes.set_xlabel('Time')
axes.set_ylabel('Occupation probability')
axes.set_title('Vacuum Rabi oscillations')
```

Out[8]: <matplotlib.text.Text at 0x7f8f0b8c3908>



Thank you!

