OPEN SCIENCE FRAMEWORK

AN OPEN SOURCE WEB APPLICATION THAT CONNECTS AND SUPPORTS THE RESEARCH WORKFLOW IN THE LIFE SCIENCES

lennart martens
lennart.martens@ugent.be
computational omics and systems biology group
Ghent University and VIB, Ghent, Belgium
Saint George and the Dragon
about 1470, Paolo Uccello

https://www.flickr.com/photos/scriptingnews/3503448168
(CC BY-SA 2.0)
We should ask ourselves a simple question: where does the real revolution lie?

Many people (myself included) have already fought long and hard for open science

But we’ve really been fighting windmills, even though some of these looked a lot like dragons

The fight may [be | have been] glorious, but in all honesty, the real challenges are:

- to make it all work in practice
- to do amazing things with open science
Many reasons for open science

An infrastructure for open science

Create a structured workspace
Create a research plan
Pre-registration
Pre-analysis plan for confirmatory research
Archive materials from study
Analyze and document analyses
Share study data, code, materials

A final call to action
Many reasons for open science

An infrastructure for open science

- Create a structured workspace
- Create a research plan
  - Pre-registration
  - Pre-analysis plan for confirmatory research
- Archive materials from study
- Analyze and document analyses
- Share study data, code, materials

A final call to action
Open science is frequently required by an important third party

Where are the data?

*Nature Biotechnology* now requires data availability statements to be supplied with research papers.

As the research community embraces data sharing, academic journals can do their bit to help. Starting this month, all research papers published in *Nature Biotechnology, Nature* and 11 other Nature titles will include information on whether and how others can access the underlying data.

These statements will report the availability of the ‘minimal data set’ necessary to interpret, replicate and build on the findings reported in the paper. Where applicable, they will include details about publicly archived data sets that have been analyzed or generated during the study. Where restrictions on access are in place—for example, in the case of privacy limitations or third-party control—authors will be expected to make this clear.

The new policy ([http://go.nature.com/2bfivqn](http://go.nature.com/2bfivqn)) builds on our long-standing support for data availability as a condition of publication. It also extends our support for data citation, the practice of citing data sets in reference lists via digital object identifiers (DOIs).
The absence of openness is increasingly considered a red flag for scientific fraud

After an investigation, the Central Ethical Review Board in Sweden has recommended the retraction of the Report “Environmentally relevant concentrations of microplastic particles influence larval fish ecology,” by Oona M. Lönnstedt and Peter Eklöv, published in Science on 3 June 2016 (1). Science ran an Editorial Expression of Concern regarding the Report on 1 December 2016 (2). The Review Board’s report, dated 21 April 2017, cited the following reasons for their recommendation: (i) lack of ethical approval for the experiments; (ii) absence of original data for the experiments reported in the paper; (iii) widespread lack of clarity concerning how the experiments were conducted. Although the authors have told Science that they disagree with elements of the Board’s report, and although Uppsala University has not yet concluded its own investigation, the weight of evidence is that the paper should now be retracted. In light of the Board’s recommendation and a 28 April 2017 request from the authors to retract the paper, Science is retracting the paper in full.
We usually think we need open science to prevent bad things from happening

While open science helps prevent some cases of fraud or low quality work being published, it is certainly not a panacea (cfr. peer review)

Simultaneously, fraud is regularly detected:

• in the absence of the source data
• from papers published in closed access journals
• without any of the code or metadata available

Why should we define the use of open science through an application with negative connotation?
Instead, we should rather focus on the good that comes from open science

Open science makes the work accessible to anyone

Open science allows people to build much more efficiently on previous work

Open science helps maximize the usefulness of each individual research effort

Data tend to have a (much!) longer shelf life than our (limited) interpretations

Open science fosters creativity, and stimulates revolutionary research
Of course, data management also forms the keystone of the data life cycle.
Open science is enabling critical work in (medical) life sciences

This work is made possible by the open sharing of genetic data by research groups from all over the world.

Not everyone agrees with data sharing
Moratoria on data reuse have been proposed.
Many reasons for open science

**An infrastructure for open science**

- Create a structured workspace
- Create a research plan
- Pre-registration
- Pre-analysis plan for confirmatory research
- Archive materials from study
- Analyze and document analyses
- Share study data, code, materials

**A final call to action**
A field guide to open science for the newly initiated

Do you speak open science? Resources and tips to learn the language

Science and Medical Education

Paola Masuzzo1,2, Lennart Martens1,2
January 3, 2017

Author and article information

Abstract

The Internet era, large-scale computing and storage resources, mobile devices, social media, and their high uptake among different groups of people, have all deeply changed the way knowledge is created, communicated, and further deployed. These advances have enabled a radical transformation of the practice of science, which is now more open, more global and collaborative, and closer to society than ever. Open science has therefore become an increasingly important topic. Moreover, as open science is actively pursued by several high-profile funders and institutions, it has fast become a crucial matter to all researchers. However, because this widespread interest in open science has emerged relatively recently, its definition and implementation are constantly shifting and evolving, sometimes leaving researchers in doubt about how to adopt open science, and which are the best practices to follow.
FAIR principles are commonly adopted, and compliant systems exist.
The Center for Open Science (COS) is the provider for the Open Science Framework (OSF)

https://osf.io/institutions/ugent
Many reasons for open science

An infrastructure for open science

Create a structured workspace

Create a research plan

Pre-registration

Pre-analysis plan for confirmatory research

Archive materials from study

Analyze and document analyses

Share study data, code, materials

A final call to action
Creating an UGent associated account on OSF

For NEW OSF users
1. Select the "Sign Up" option at http://osf.io
2. Select the "Login through your Institution" option
3. Select “Universiteit Gent” from the list of available institutions
4. User will be directed to complete their account creation
5. Email verification will be sent
5. Login to the OSF using the link provided

For EXISTING OSF users
1. Go to https://accounts.osf.io/login?campaign=institution (or select "Login through your institution" from the main login page)
2. Select “Universiteit Gent” from the list of available institutions
3. If you have an account on the OSF with your UGent email address ALREADY associated with your account, the step above will verify your credentials and bring you back to a verification screen
4. Users can then login into the OSF via the regular sign in process. The institutional sign in step only needs to be completed once by each user
Once UGent is associated with your login, you can make your projects UGent affiliated.

Flag your project as UGent through the project settings.

This association is shown by the UGent logo on the project.

As soon as you make a Ugent-associated project public, it will appear on the UGent landing page of the OSF:

https://osf.io/institutions/ugent
DEMONSTRATION
Many reasons for open science

An infrastructure for open science

Create a structured workspace

**Create a research plan**

Pre-registration

Pre-analysis plan for confirmatory research

Archive materials from study

Analyze and document analyses

Share study data, code, materials

A final call to action
Pre-registration entails the detailing of the study \textit{before} you begin the study

General information about what you are investigating, and how you intend to do this:

- Research question
- Population and sample size
- General study design
- Variables you’ll be collecting, or data set you’ll be using

Allows discovery of unpublished, \textit{null} or unexplored findings
Positive results are heavily overrepresented in the literature, biasing our view.
Many reasons for open science

An infrastructure for open science
  Create a structured workspace
    Create a research plan
      Pre-registration
        Pre-analysis plan for confirmatory research
    Archive materials from study
    Analyze and document analyses
    Share study data, code, materials

A final call to action
A pre-analysis plan details the analyses you will run, and thus restricts your freedom

You have to detail your planned confirmatory analyses in advance

If it seems as if this limits your freedom: that is the point!

So-called ‘researcher degrees of freedom’ are dangerous, and are linked to ‘p-value hacking’

So before you start, register in read-only format:

- Sample size
- Data processing and cleaning procedures
- Exclusion criteria
- Statistical analyses to be performed
A whimsical but absolutely apt example

From: http://fivethirtyeight.com/features/science-isnt-broken/#part1
There is a continuum ranging from exploratory to confirmatory studies

Wholly exploratory study can make do with pre-registration of data collection only

Partially exploratory, partially confirmatory study should provide pre-registration for all data collection, as well as a pre-analysis plan for the confirmatory analyses

A wholly confirmatory study needs to have a pre-registration of the data collection, and a pre-analysis plan for the whole study
Many reasons for open science

An infrastructure for open science

Create a structured workspace
Create a research plan
Pre-registration
Pre-analysis plan for confirmatory research

**Archive materials from study**
Analyze and document analyses
Share study data, code, materials

A final call to action
Data management requires 3 building blocks: minimal requirements, CVs, and formats
Metadata is often the key issue, as it requires the most manual work.
Even user-friendly submission tools cannot correct for a lack of elementary motivation.
Manual curation of submissions, equivalent to restrictive policing, does help
A non-exhaustive checklist to ensure suitability of your data archival efforts

Are all abbreviations explained?

Are all metadata properly annotated?

Are standard formats used to represent the data in?

Are all terms derived from a controlled vocabulary?

Are missing data explained?

Am I using a trusted, long-lived, third-party repository?
The available links to third-party content hosting services from the OSF
Total planned and in-progress links to third-party hosting content services
Many reasons for open science

An infrastructure for open science

Create a structured workspace
Create a research plan
  Pre-registration
  Pre-analysis plan for confirmatory research
Archive materials from study

**Analyze and document analyses**

Share study data, code, materials

A final call to action
When it comes to the analysis of your data, your paper contains the advertisement...

Gross, Nature Genetics, 2014
... but the code on GitHub represents the actual research performed

[GitHub page with code snippets]

https://github.com/theandygross/TCGA
Code should also always come with an appropriate license, especially open code.
Many reasons for open science

An infrastructure for open science

Create a structured workspace
Create a research plan
Pre-registration
Pre-analysis plan for confirmatory research
Archive materials from study
Analyze and document analyses

Share study data, code, materials

A final call to action
Non-exhaustive checklist for sharing (sensitive) data with the world

Do I have the proper clearances from study participants and/or other data originators?

Double-check clearances for: sharing, reuse, combination

Is the data properly anonymized/de-identified?

These things should really have already been sorted up front, with the ethical committee and other relevant experts!

Does the data have a well-considered license?
Creative Commons (CC) licenses for any content that is not software code

Data without license may NOT be shared at all

Two Creative Commons licenses should be your top choices:

- Attribution (CC BY), only mandates recognition of the author
- Attribution-ShareAlike (CC BY-SA), as CC BY above, but all derived works need to be licensed CC-BY-SA as well (infectious licence)

Note that CC licenses are not meant for software code; see earlier for examples of suitable open source licenses!

Wikipedia and a lot of Flicr use CC, amongst many others

https://creativecommons.org; https://creativecommons.org/licenses
Many reasons for open science

An infrastructure for open science

Create a structured workspace
Create a research plan
  Pre-registration
  Pre-analysis plan for confirmatory research
Archive materials from study
Analyze and document analyses
Share study data, code, materials

A final call to action
Here is treasure of unlimited size, with all dragons chased away – will you contribute?

https://www.flickr.com/photos/fantasy-art-and-portraits/2884954207
(CC BY-NC-SA 2.0)