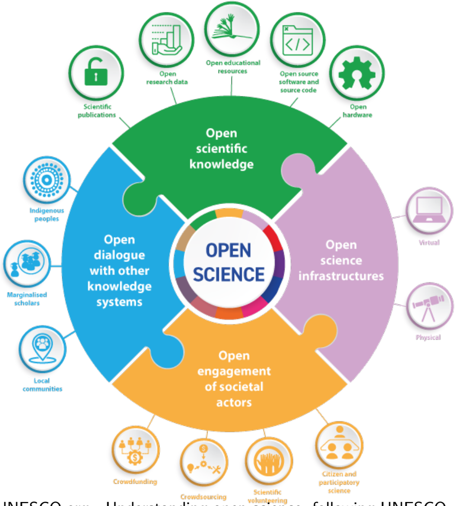
Introduction to open-science and reproducibility in scientific computing

Céline Acary-Robert (Grenoble Alpes University)
LJK, GRICAD

Open Science

Unesco definition



By UNESCO.org - Understanding open science, following UNESCO Recommendation on Open Science, CC BY-SA 4.0.

Open science is the movement to make scientific research transparent and accessible to all levels of society through collaborative networks including:

- publications
- data
- physical samples
- software

Open Science

Openness in research

Each element of the research process should be:

- publicly available: easy to find and to use/ without barriers
- transparent: with appropriate metadata to provide clear statements of how research output was produced and what it contains.
- reusable: research outputs need to be licensed appropriately

Open Science

What benefits will the research bring?

- Increase the speed of knowledge dissemination
- Accelerate scientific progress
- Societal and economic benefits: open science supports innovation, education, ...
- Society and general public are more confident in research

Reproducibility: main concept

Context: a reader/reviewer of a publication would like to reuse some artifacts and:

- reproduce an existing result or change a parameter value
- extend some functionalities of the software

He needs:

- the article
- the code
- the associated dataset

→ Open Science principles is a necessary step to insure reproducibility

Reproducibility

Benefits: what is the interest for you ?

Examples:

- → you have a suggestion from a reader or colleague to improve something
- → you want to reuse your code two years later ...
- → a phD student want to compare her/his own results with a former study

If your results are reproducible, it means that:

- you can quickly reuse your code later
- At each stage of the development process, you can reuse your code and trust the specific version.
- you save (a lot of) time later

Reproducibility

What benefits will the research bring?

- Make research results sustainable
- Build trust with readers/reviewers
- Increase research efficiency by reusing existing work
- Improve code quality: bugs can be reported
- Promote collaboration: contributions can be made, etc.

➜ Increase confidence in research results for you, your colleagues and also the general public

Open Science
Main Objectives

- to disseminate datasets, codes and publications, without barrier
- to cumulate knowledge
- to encourage collaborations
- to foster scientific integrity

Open Science: key dates

- 2002 – Budapest Open Access Initiative: foundation of the open access movement.
- 2003 – Berlin Declaration: cornerstone of open science in Europe.
- 2016 – Launch of the European Open Science Cloud (EOSC).
- 2018 – Plan S: mandatory immediate open access for publicly funded research.
- 2021 – Horizon Europe: open science becomes a core principle of EU research policy

Open Science: international framework

- Center for Open Science (COS, USA) https://www.cos.io/
  ➜ infrastructures (OSF), training and metaresearch
- European projects to promote Open Science (openAIRE, ...)
- European Open Science Cloud (EOSC) https://eosc.eu/
  ➜ federated infrastructure
  ➜ aims to contribute to a "web of FAIR data and services" for science in Europe
- Global Research Initiative on Open Science (GRIOS) https://www.grios.org/
  ➜ aims to analyze the impact of open science policies in research and increase the efficiency of their open science strategies.
- ...

Open Science in France

➜ Laws for Open Science and integrity (Lois pour une république numérique)

Organization

- COSO: COmité pour la Science Ouverte: French Ministry of Research
- Local data management units: "Ateliers de la donnée"

Objectives

- Generalize open access to publications
- Structure, share and open data from public research
  ➜ promote FAIR data
- Promote open source codes/software from public research
- Transform research practices to make open science the standard

Open Science: FAIR definition

- FAIR data: Findable, Accessible, Interoperable and Reusable
  - Findable:
    - in a public repository, with a unique and perennial identifier (DOI)
    - with rich and standard metadata for an accurate description
  - Accessible
    - with description of access conditions
    - as free as possible
  - Interoperable
    - using open format
    - providing code source to read or analyze data
  - Reusable
    - with a license
    - with documentation to facilitate their use
- With restrictions if any: as open as possible, as closed as necessary
  private company, medical data in research, defense/army, personal data

Reference: https://www.nature.com/articles/sdata201618
FAIR by design: methodological guide

Open Science: how to ?

- FAIR code:
  - Findable:
    - on a public forge
    - with rich and standard metadata for a precise description
  - Accessible
    - with description of access conditions
    - as open as possible
  - Interoperable [?]
  - Reusable
    - with an open license
    - with documentation to facilitate its use

Maybe CODE for Open, Document, Execute and Collaborate
Roberto Di Cosmo, Sabrina Granger, Konrad Hinsen, Nicolas Jullien, Daniel Le Berre, et al.. CODE beyond FAIR. 2025.
https://hal.science/hal-04930405

Open Science: how to ?

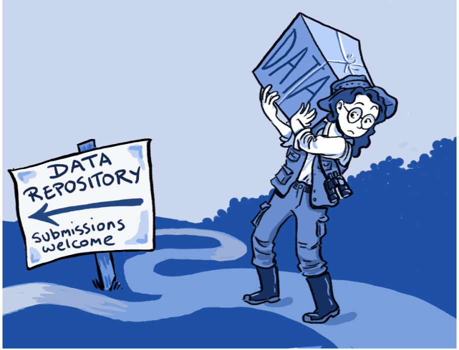
- Dataset: available on a repository with a perennial identifier + metadata + license + documentation
- Code: available on a public forge with a perennial identifier + license + documentation
- Article: in an open repository: HAL, ArXiv, etc ...

➜ 💡 Remark: repositories with institutionnal endorsement to insure reliability and sustainability

➜ passport-for-open-science

Open Science
Barriers

- lack of awareness and training
- takes time and requires extra effort
- not enough considerated for carreer and promotion
- fear of being judged by other researchers
- fear of reuse before publication (embargo)



By Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) - Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, et al. (2014) CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=30978545

Open Science

Benefits

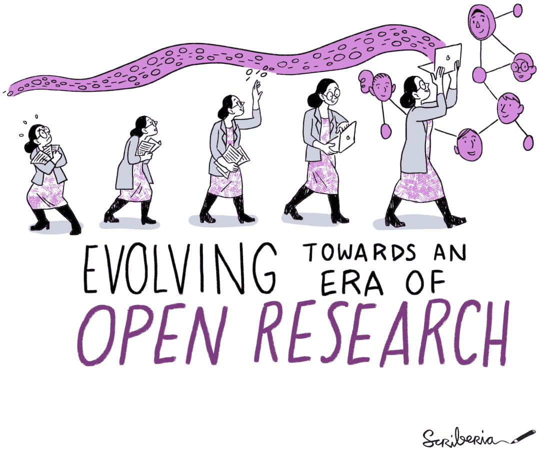
- access, transparency, collaborations
- open source practices increase citation:

  How open science helps researchers succeed, E. McKiernan, 2016
  We review literature demonstrating that open research is associated with increases in citations, media attention, potential collaborators, job opportunities and funding opportunities.
  https://elifesciences.org/articles/16800

  An analysis of the effects of open science indicators on citations in the French Open Science Monitor Giovanni Colavizza, Lauren Cadwallader, Iain Hrynaszkiewicz, 2025
  This study investigates the correlation of citation impact with various open science indicators (OSI) within the French Open Science Monitor.
  https://doi.org/10.48550/arXiv.2508.2074

- Pilar for reproducibility

Open Science



The Turing way from the Alan Turing Institute, UK institute for data science and AI

Why reproducibility matters ?

Fundamental principle: scientific results have to be tested and validated by others researchers to become "a knowledge", need a consensus

- Ensures reliability: helps to verify that conclusions are not due to errors
- Optimizes resource use: better use of time, funding, and infrastructure
- Accelerates scientific progress: avoids unnecessary duplication of experiments
- Enables reuse: data, code, and methods can be reused by other researchers
- Improves research quality: encourages rigorous and transparent practices
- Increases trust: reproducible results strengthen confidence in scientific findings.

Historical context

Various studies about the reproducibility of results

- Researchers started to study the reproducibility of published results in 2000's
- Studies conducted as part of research on research methods: metascience, metaresearch:
  - interdisciplinary field
  - aims to analyze, evaluate, and understand the scientific method, research practices
- A lack of reproducibility in many publications identified: "reproducibility crisis", 2010's

Lack of reproducibility

Main factors

- data or code not available or not usable
- selective reporting
- publication race: lack of time
- lack of awareness, training, support
- human errors in publication
- ...
- rare cases: fraud

Survey conducted among 1500 scientists

Baker M. 2016, 1,500 scientists lift the lid on reproducibility, Nature 533:452454

https://doi.org/10.1038/533452a

Study of reproducibility

- Economy
  - Study of the reproducibility conducted over 60 articles (with available data + code + pre-analysis)
  - 1/3 articles: reproduced, +10% after contact with authors to get more details

    Chang, Andrew C., and Phillip Li. 2017. "A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time." American Economic Review, 107 (5): 6064. DOI: 10.1257/aer.p20171034
- Applied psychology
  - Study of the reproducibility of the results conducted over 35 articles (published between 2015 and 2018 in the review *Cognition* with available data
  - 11 results in articles reproducible, 11 more after contact with authors: 63% max

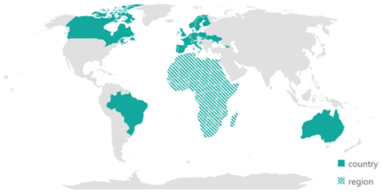
    Estimating the Prevalence of Transparency and Reproducibility-Related Research Practices in Psychology (20142017), Tom E. Hardwicke, Robert T. Thibault, [], and John P. A. Ioannidis, Volume 17, Issue 1, https://doi.org/10.1177/1745691620979806
- Ten Years Reproducibility Challenge Oct 11, 2019
  - Main objective: rerun a published code ten years later
  - Difficulties: find the code and set up the same software environement !

    Challenge to scientists: does your ten-year-old code still run? Nature vol. 584 iss. 7822, 2020

International context

Emergence of structuration

- This is a matter of concern within various scientific communities
- Some initiatives have emerged at a national level:
  - establishment of national reproducibility networks
  - various structurations



International context

- Global Federation of national networks GFRN
- European projects: horizon europe (european program): some projects around metascience and reproducibility are funded (OSIRIS, iRISE, TRUSTparency, ...)

Objectives: to pool efforts to

- raise awareness of the issue of reproducibility in various scientific fields
- foster best pratices providing infrastructures and tools
- organize conferences and workshops
- share training material

French Reproducible Research Network

- created in 2023
- informal structure
- around 400 members
- support of the French Ministry with a project manager

Liste de diffusion

Main activities

- awareness webinars to share practices
- workshops
- bibliographic database
- working group around software to write “best practices checklist”
- ...

https://www.recherche-reproductible.fr

Reproducibility

A great diversity

Good practices will depend on:

- scientific discipline
- implemented technics
- data management
- etc ...

➜ Strong need to fix the terminology at an international level

Definitions

Different kinds of reproducibility

- empirical: record of the conditions under which the experiment took place
- observational: record of the conditions under which the observations were made
- statistical: evidence of the reason why a result is statistically significant
- computational: record of the production of a result on a computer

Definitions

Different kinds of reproducibility (1)

- empirical: record of the conditions under which the experiment took place (ex: chemistry)
    - parameters: temperature, pH, etc ,
    - differents elements in the experiment: description of the chemical products
    - implementation protocol (steps, material, etc)
    - ...

Definitions

Different kinds of reproducibility (1)

- empirical: record of the conditions under which the experiment took place (ex: chemistry)
  - parameters: temperature, pH, etc ,
  - differents elements in the experiment: description of the chemical products
  - implementation protocol (steps, material, etc)
  - ...
- observational: record of the conditions under which the observations were made (ex: biodiversity study
  - observation protocol
  - context, duration, conditions of the observation
  - list of the observed elements, evaluation criterion, etc.
  - different steps of the observation, area, timetable, tools, etc ...

Definitions

Different kinds of reproducibility (1)

- empirical: record of the conditions under which the experiment took place (ex: chemistry)
  - parameters: temperature, pH, etc ,
  - differents elements in the experiment: description of the chemical products
  - implementation protocol (steps, material, etc)
  - ...
- observational: record of the conditions under which the observations were made (ex: biodiversity study
  - observation protocol
  - context, duration, conditions of the observation
  - list of the observed elements, evaluation criterion, etc.
  - different steps of the observation, area, timetable, tools, etc ...

➜ experimental or observational conditions may be very hard to recreate !

Definitions

Different kinds of reproducibility (2)

- statistical: evidence of the reason why a result is statistically significant
  - justification for the choice of statistical test
  - choice of the model parameters
  - threshold values, sample sizes
  - ...

Definitions

Different kinds of reproducibility (2)

- statistical: evidence of the reason why a result is statistically significant
  - justification for the choice of statistical test
  - choice of the model parameters
  - threshold values, sample sizes
  - ...
- computational: record of the production of a result on a computer
  - code + version + workflow, parameters file
  - hardware, software environment, dependancies
  - datasets
  - ...

➜ nowadays computational aspects are very transdisciplinary

Repeatability, Reproducibility and Replicability

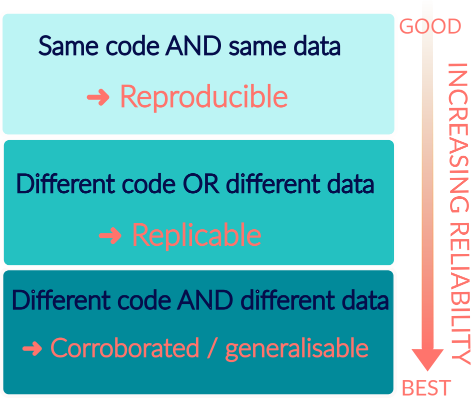
Preliminary remarks:

- Definitions depend on the scientific fields: a need for standardization
- a need to define in a specific scientific field what mean same results or equivalent conclusions/results

Definitions for scientific computing:

- Repeatability: same team, same code/protocol and same data lead to same *results*
- Reproducibility: different team, same code/protocol and same data lead to same *results*
- Replicability: different team, different code/protocol or different data lead to same *results*

Reliability



Sources of non-reproducibility in scientific computing

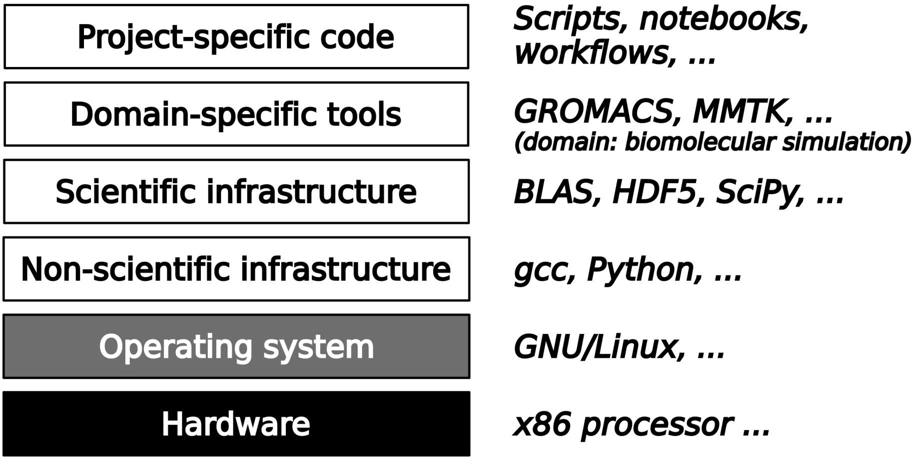
→ We suppose that datasets and code are available on a repository but:

- you are not able to compile or run the code
- or you get different results

→ Reasons:

- Difference between version of the compiler / interpreter or dependances
- Optimization options at the compilation stage
- Roundind errors accumulation
- parallelism, changing operation orders
- CPU / GPU
- ...

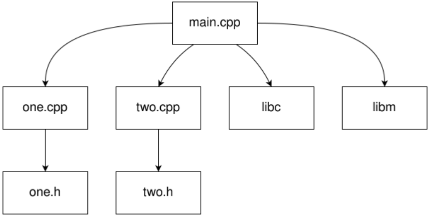
Dependencies management



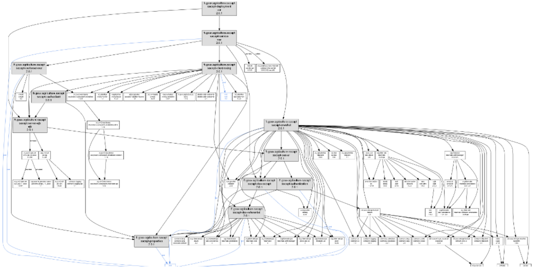
Konrad Hinsen, https://blog.khinsen.net/

Dependencies management

code



dependancies



Examples of reproducibility issues in scientific computing

Version of python3 :

With Python 3.6, default values of rel_tol and abs_tol have been modified:

python3.5 import math math.isclose(1.0, 1.0000001) :: False	python3.6 import math math.isclose(1.0, 1.0000001) :: True

Examples of reproducibility issues in scientific computing

Version of dependancies: syntax evolution

NumPy $<$ 2.0 **removed member**	NumPy $>=$ 2.0 **migration guideline**
Inf	Use np.inf instead
mat	Use np.asmatrix instead
NaN	Use np.nan instead
longfloat	Use np.longdouble instead

https://numpy.org/doc/stable/numpy_2_0_migration_guide.html

Pandas $<$ 0.20	Pandas $>=$ 0.20
df.sort('column_name')	df.sort_values('column_name')

Examples of reproducibility issues in scientific computing

Compilation, parallelization

$$H_n = \sum_{k=1}^{n} 1/k = ln(n) + \gamma + o(1)$$

- Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz
- gfortran 14 -fopenmp

naive algorithm with double precision and OpenMP parallelisation:

```
Gamma = 0.5772156649015328606
!$OMP parallel do private(i) reduction(+ : sum)
do i=1,n
   sum = sum + 1.0/i
enddo
!$OMP end parallel do
p = abs(sum - log(float(n)) - Gamma)
```

Examples of reproducibility issues: OpenMP parallelisation

	n= 1 000 000	n= 1 000 000 000
Sequentiel	14.392726788474306	21.300481573265063
OpenMP 1 thread	14.392726788474306	21.300481573265063
OpenMP 4 threads	14.392726788474306	21.300481572647087
OpenMP 4 threads -Ofast	14.392726222394913	21.300480911325508

Examples of reproducibility issues in scientific computing: GPU

- Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz
- 1 GPU nvidia V100
- nvfortran 22.3-0 64-bit target on x86-64 Linux

```
Gamma = 0.577215664901532860 6
!$acc kernels copy(sum)
devicetype = acc_get_device_type()
ngpus = acc_get_num_devices(devicetype)
do i=1,n
   sum = sum + 1.0/i
enddo
!$acc end kernels
p = abs(sum - log(float(n)) - Gamma)
```

nvfortran

n= 1 000 000

Calcul	CPU	GPU
sum	14.39272678847431	14.39272678847431

n= 1 000 000 000

Calcul	CPU	GPU
sum	21.30048157326506	21.30048156797372

➜ what level of numerical precision is significant ?

Computational reproducibility

Good practices

- Interpreter version and dependancies
  - ➜ share software environment
  - ➜ maintain the code over time
- Precision
  - ➜ choose the significant precision you need
  - ➜ share compilation and execution details

Scientific computing: good practices

Software developpement

- Implement best practices for development
  ➜ to improve its portability and reusability
- Use a reproducibility-oriented software manager (such as Guix, Nix, pixi ...) to improve reproducibility
- Think carefully about a license: this will regulate the use of the code
- Use a versioning tool such as git for example
- Write a documentation
- Use a software forge to automate tests and documentation

Scientific computing: good practices

Software dissemination

The repository should contain:

- your code (!)
- authors and contributors information
- a license file
- a brief description of the code in a README file
- a brief description of the worflow to use it
- a documentation (users and contributors)

Scientific computing: good practices

Software execution

The repository should contain:

- description of the hardware
- software environment
  - precise description of the software environment
  - the way to re-create it
- input parameters file
- workflow for execution

➜ For each step there are some useful tools !

Scientific computing

Software in publications: how to cite?

Software Heritage
https://www.softwareheritage.org/

Software Heritage main objectives are to:

- Collect and preserve software in source code form
- Curate and make accessible all the collected software
- Preserve in the very long term.
- Provide a persistant identifier: Software Heritage ID (SWHID)
  https://archive.softwareheritage.org/save/

Scientific computing

Summary

➜ In the article:

- Share the datasets (input and output): DOI
- Share the code: provide SWHID of the exact version

➜ Share your article in Open Archive

Nobody is perfect !

Key point: slowly change and gradually adopt good practices 😉

- Think about your own practices to identify obstacles
  - technical obstacles
  - structural obstacles: pressure to publish
  - etc ...
- Identify points that can be improved
- Identify the tools to achieve this improvment
- Learn gradually how to use them

Conclusion

From a practical point of view

- At the beginning, it requires time ....
- It's a gradual process
- But we save a lot of time later

Who is the winner?

- You: rigor, time, confidency, ...
- Reviewers: save time and confidency
- Your collegues: share your code with them is easy, collaborations are easy
- Public in general: confidency in science

Some bibliographical references

- MOOC reproducible research (funmooc platform)
  - https://www.fun-mooc.fr/en/courses/reproducible-research-methodological-principles-transparent-scie/
  - https://www.fun-mooc.fr/fr/cours/reproducible-research-ii-practices-and-tools-for-managing-comput/
- https://inria.hal.science/hal-01872189v2/
- https://indico.math.cnrs.fr/event/10998/contributions/10948/attachments/4774/7400/SWH-HAL_RNBM~Mathrice~v7.pdf
- https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf
- https://inria.hal.science/hal-04930405v1/document
- NicolasBonneel,DavidCoeurjolly,JulieDigne,NicolasMellado,CodeReplicabilityinComputerGraphics, ACMTrans.onGraphics(ProceedingsofSIGGRAPH2020),39:4
- https://www.acm.org/publications/artifacts
- https://www.sciencedirect.com/science/article/abs/pii/S157401372400039X

Thank you

Enjoy open science and reproducibility

😉