Why we need FAIRness

University of Pisa - Phd course 2022

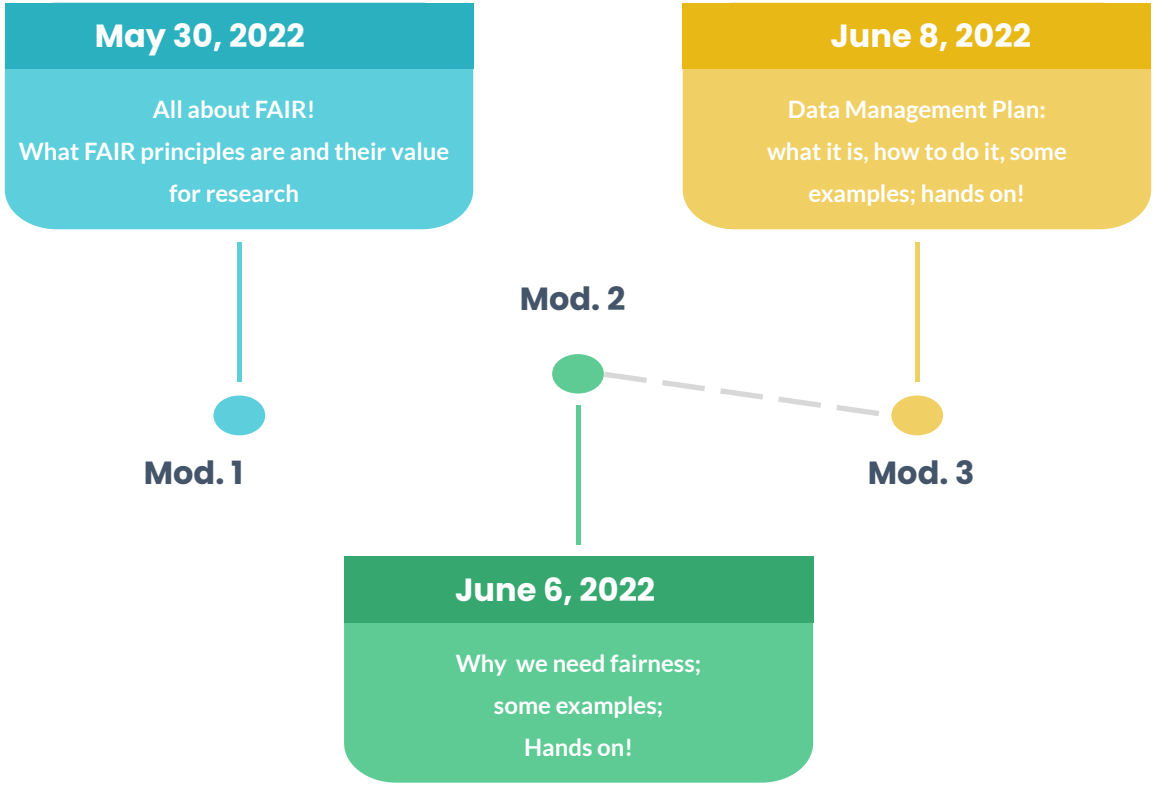
Gina Pavone, CNR 0000-0003-0087-2151

Day 2 - 6 June 2022

10.5281/zenodo.6597239

FAIR data and DMP

COURSE OUTLINE



Today's agenda

University of Pisa
PhD students

A quick recap

FAIR data infographic (CC-BY except F.A.I.R logos CC-BY-SA by Sangya Pundir

Accessible doesn’t mean Open

Software sharing: PIDs are still needed

https://guides.github.com/activities/citable-code/

Data are not yours

Data is **not** intellectual work, it is fact and information

Copyright protection covers expressions and not ideas, procedures, operating methods or mathematical concepts as such.

Protection is on databases and not on data. Data are protected only and especially when they are collected and organized in a database.

The sui generis property right (**only in Europe**) covers not only the reproduction and dissemination of the database, but also the extraction and reuse of substantial parts of the database.

A valid resource (in Italian)

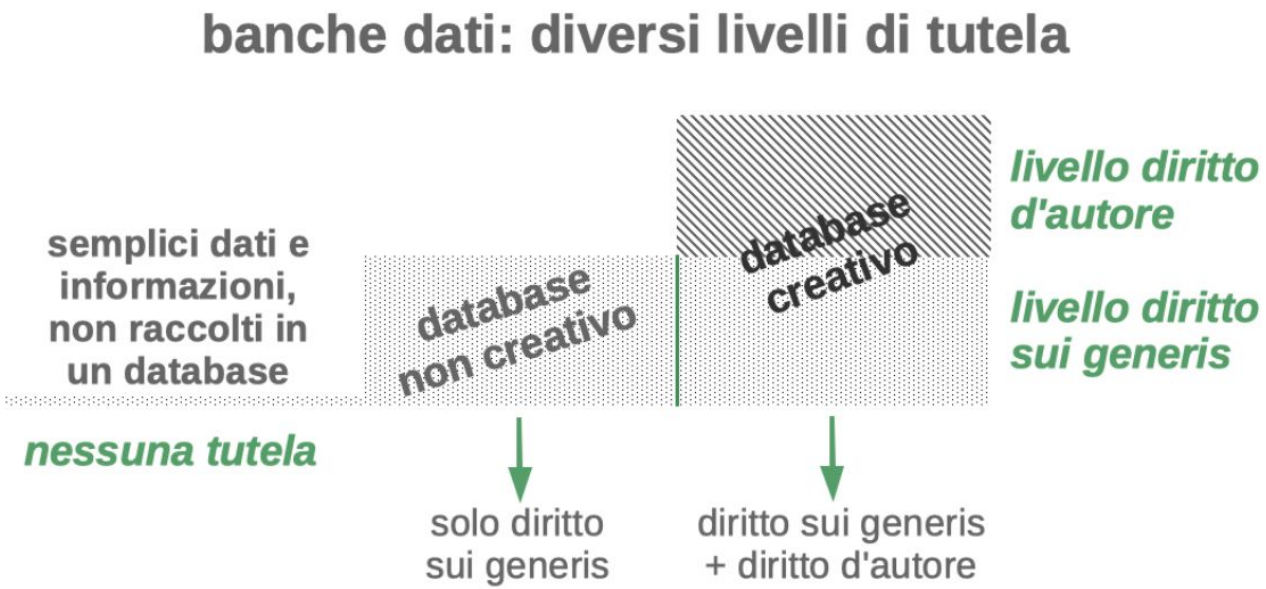


Figura 1: I due diversi livelli di tutela per le banche dati basati sul requisito del carattere creativo

Aliprandi, Simone. (2022). Aspetti legali degli open data: la guida definitiva (1.0 (maggio 2022)). Zenodo. https://doi.org/10.5281/zenodo.6575822

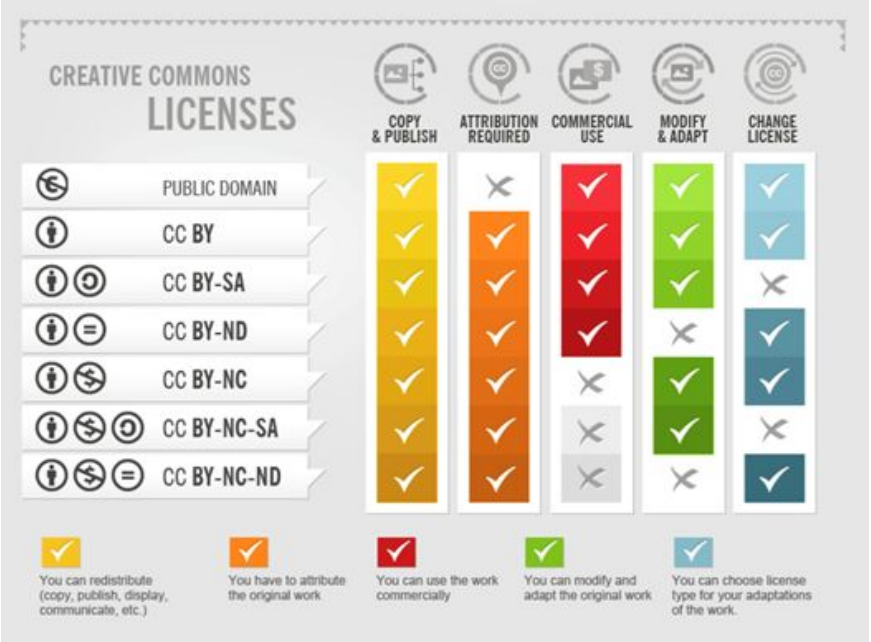

Licenses
Tell other what they
can do with your data

Creative Commons

Not all of us are legal experts capable of writing proper licenses.

Creative Commons and Public Domain create legal certainty for everyone, who wants to use works, that are licensed respectively.

It is important to follow and understand the different meanings of the licenses and follow the rules for using them.



Why FAIR

Making your research data Findable, Accessible, Interoperable and Reusable could:

- Help **peers** and your **future-self** understand the research project and data
- Facilitate data sharing and **collaborations**
- Increase the visibility of research and can lead to more **citations**
- Improve the **transparency, reliability and reproducibility** of research
- Prevent data loss

And thereby:

- Maximise potential from data assets
- Maximise research **impact**

https://www.howtofair.dk/why-fair/

Not only publications

Science is also:

- data
- software
- protocols
- negative results
- lab notes
- project deliverable
- and much more…

26 billion €
are going lost every year in Europe for not managing the data properly

Directorate-General for Research and Innovation (European Commission), PwC EU Services **Cost of not having FAIR research data**, 2019-01-16

“Publishing research without data is simply advertising, not science”

Graham Steel

What happens if...

We do not manage and share research data in the correct way?

Data can be lost...

JAMA journal retracts paper when author can't produce original data

In July 2017, a *JAMA* journal called for an investigation into a 2013 paper it had published after concluding that the article had "scientific and ethical concerns." Now the journal, *JAMA Otolaryngology – Head & Neck Surgery*, is retracting the paper.

The article, "Dexamethasone for the prevention of recurrent laryngeal nerve palsy and other complications after thyroid surgery: a randomized double-blind placebo-controlled trial," came from a group in Italy led by Mario Schietroma, of the Department of Surgery at the University of L'Aquila, in Abruzzo, Italy. Schietroma, who in December admitted to us that a retracted 2015 paper of his in the *Journal of the American College of Surgeons* suffered from "misinterpretation of the statistical data," now has four retractions.

Neither [the original dataset and the approved protocol] have been provided by Dr Schietroma, and the university has informed us that ***"without those pieces of information the results of the papers under investigation cannot be validated."***

https://retractionwatch.com/2018/10/25/jama-journal-retracts-paper-when-author-cant-produce-original-data/

The importance of depositing research data

MENU nature

Carlisle has kept going. This year, he warned about dozens of anaesthesia studies by an Italian surgeon, Mario Schietroma at the University of L'Aquila in central Italy, saying that they were not a reliable basis for clinical practice[6]. Myles, who worked on the report with Carlisle, had raised the alarm last year after spotting suspicious similarities in the raw data for control and patient groups in five of Schietroma's papers.

Bottled oxygen, used by anaesthetists during surgery. Credit: Mark Thomas/Alamy

The challenges to Schietroma's claims have had an impact in hospitals around the globe. The World Health Organization (WHO) cited Schietroma's work when, in 2016, it issued a recommendation that anaesthetists should routinely boost the oxygen levels they deliver to patients during and after surgery, to help reduce infection. That was a controversial call: anaesthetists know that in some procedures, too much oxygen can be associated with an increased risk of complications – and the recommendations would have meant hospitals in poorer countries spending more of their budgets on expensive bottled oxygen, Myles says.

The five papers Myles warned about were quickly retracted, and the WHO revised its recommendation from 'strong' to 'conditional', meaning that clinicians have more freedom to make different choices for various patients. Schietroma says his calculations were assessed by an independent statistician and through peer review, and that he purposely selected similar groups of patients, so it's not surprising if the data closely match. He also says he lost raw data and documents related to the trials when L'Aquila was struck by an earthquake in 2009. A spokesperson for the university says it has left enquiries to "the competent investigating bodies", but did not explain which bodies those were or whether any investigations were under way.

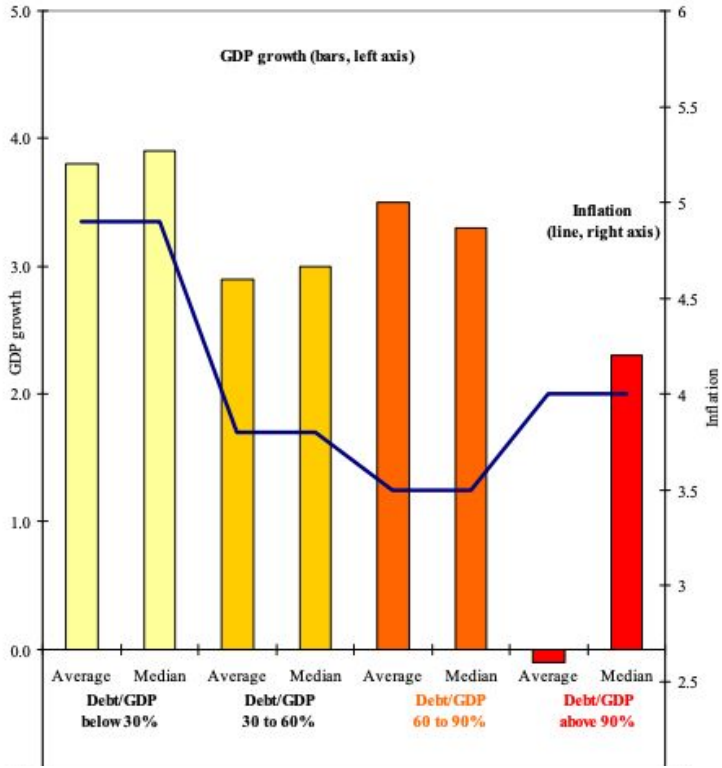
https://www.nature.com/articles/d41586-019-02241-z

Errors can get undetected

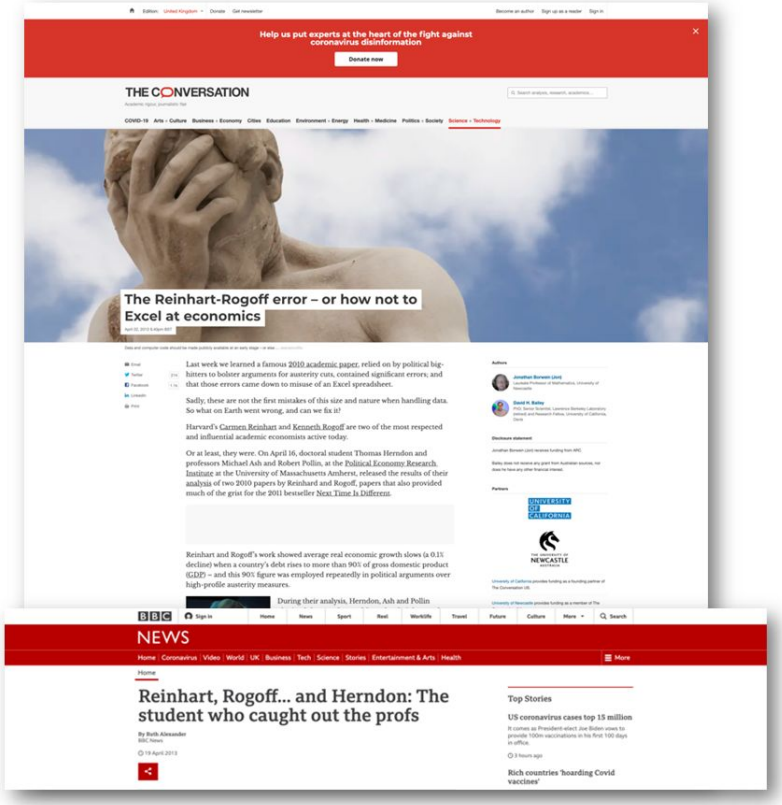
The case with austerity theory.

- The thesis: economic growth slows down dramatically when the size of a country's debt exceeds 90% of gross domestic product.
- The results shown in the paper were used to support public austerity policies during the recent economic crisis.
- But some considerations were based on wrong calculations.
- This was discovered by a student who could not replicate the results and asked the authors for the original dataset.

Figure 2. Government Debt, Growth, and Inflation: Selected Advanced Economies, 1946-2009



Errors and miscalculations



Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon* Michael Ash Robert Pollin

April 15, 2013

Herndon, 2013

JEL CODES: E60, E62, E65

Abstract

We replicate Reinhart and Rogoff (2010a and 2010b) and find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period. Our finding is that when properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of over 90 percent is actually 2.2 percent, not −0.1 percent as published in Reinhart and Rogoff. That is, contrary to RR, average GDP growth at public debt/GDP ratios over 90 percent is not dramatically different than when debt/GDP ratios are lower.

We also show how the relationship between public debt and GDP growth varies significantly by time period and country. Overall, the evidence we review contradicts Reinhart and Rogoff's claim to have identified an important stylized fact, that public debt loads greater than 90 percent of GDP consistently reduce GDP growth.

Data can be manipulated

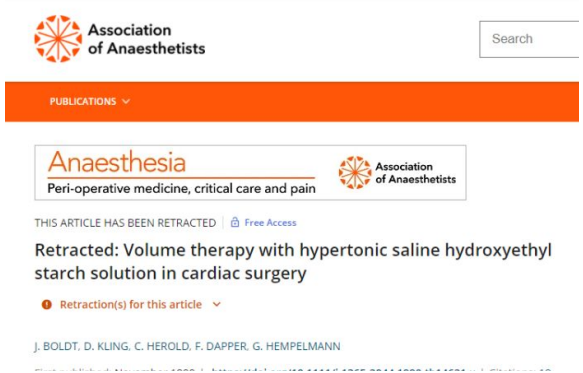
Nikolai Ivanovich Yezhov was a Soviet secret police official under Joseph Stalin who was head of the NKVD from 1936 to 1938, during the height of the Great Purge. After he fell from Stalin's favour he was executed. Among art historians, he also has the nickname "The Vanishing Commissar" because after his execution, his likeness was retouched out of an official press photo; he is among the best-known examples of the Soviet press making someone who had fallen out of favour "disappear".

The Newseum (1 September 1999). “The Commissar Vanishes" in The Vanishing Commissar”. Archived from the original on 8 February 2007.
https://en.wikipedia.org/wiki/Nikolay_Yezhov

Data can be manipulated

Largely used drug in surgery due to evidence in scientific papers.

- An investigation by Jacob-Liebig-University into the research conduct of J. Boldt.
- They found that data included in his student's MD theses were significantly different from the data in the published research papers where J. Boldt was the responsible author
- Findings in the investigation include changed mean values and standard deviations providing considerably more positive results



Anaesthesia 2021, 76, 563 doi:10.1111/anae.15378

Retraction

Retraction: Boldt J, Kling D, Herold C, Dapper F, Hempelmann G. Volume therapy with hypertonic saline hydroxyethyl starch solution in cardiac surgery. *Anaesthesia*. The above article from the Anaesthesia, published online November 1990 in Wiley Online Library (https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2044.1990.tb14621.x?sid=nlm%3Apubmed), and in Volume 45, pp. 928–34, has been retracted by agreement between the journal's Editor-in-Chief, Dr Andrew Klein and John Wiley and Sons Ltd. This retraction follows an investigation by Jacob-Liebig-University into the research conduct of J. Boldt while teaching at the university and working at the Hospital of the City of Ludwigshafen. They found that data included in his student's MD theses were significantly different from the data in the published research papers where J. Boldt was the responsible author. Findings in the investigation include changed mean values and standard deviations providing considerably more positive results . Therefore the Jacob-Liebig-University recommends that journal editors retract all papers where J. Boldt is the responsible author.

Data can be invented

REPORT

Coping with Chaos: How Disordered Contexts Promote Stereotyping and Discrimination

Diederik A. Stapel[1,*], Siegwart Lindenberg[1,2,*]

+ See all authors and affiliations

Science 08 Apr 2011:
Vol. 332, Issue 6026, pp. 251-253
DOI: 10.1126/science.1201068

Article | Figures & Data | Info & Metrics | eLetters | PDF

This article has been retracted. Please see:
Is retracted by - December 02, 2011

Abstract

Being the victim of discrimination can have serious negative health- and quality-of-life–related consequences. Yet, could being discriminated against depend on such seemingly trivial matters as garbage on the streets? In this study, we show, in two field experiments, that disordered contexts (such as litter or a broken-up sidewalk and an abandoned bicycle) indeed

- 58 articles published by Diederik Stapel were withdrawn because they were based on invented data.
- His papers had been published in scientific journals considered prestigious (very high IFs!).
- Following reports from three doctoral students, the Dutch university for which he worked had started an investigation. Stapel then admitted that he had fabricated the data on numerous occasions.
- If he had shared his data before, he probably wouldn't have been able to fabricate fakes for so long.
- This case led the Netherlands become one of the pioneer countries in Open Science policy and practices

http://www.insidehighered.com/news/2011/11/28/scholars-analyze-case-massive-research-fraud

Reproducibility crisis

80% of data are unavailable after 20 years

In their parents' attic, in boxes in the garage, or stored on now-defunct floppy disks — these are just some of the inaccessible places in which scientists have admitted to keeping their old research data.

Such practices mean that data are being lost to science at a rapid rate

https://www.cell.com/current-biology/fulltext/S0960-9822(13)01400-0
https://www.nature.com/articles/nature.2013.14416

Current Biology

Log in Register

REPORT | VOLUME 24, ISSUE 1, P94-97, JANUARY 06, 2014

PDF [462 KB] Figures

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines • Arianne Y.K. Albert • Rose L. Andrew • ... Jean-Sébastien Moore • Sébastien Renaut • Diana J. Rennison • Show all authors

Open Archive • Published: December 19, 2013 • DOI: https://doi.org/10.1016/j.cub.2013.11.014 •

Check for updates

Highlights

Summary

Results

Comments

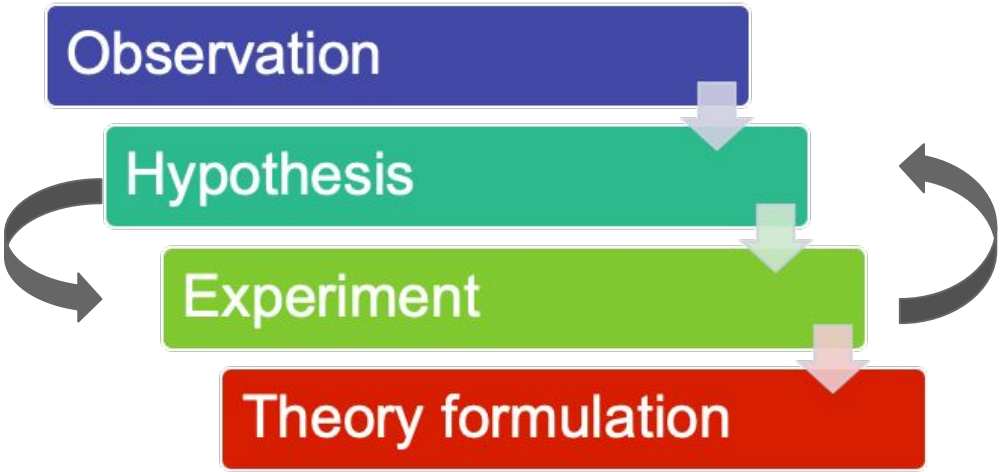
Highlights

- We examined the availability of data from 516 studies between 2 and 22 years old
- Policies mandating data archiving at publication are clearly needed

Summary

Policies ensuring that research data are available on public archives are increasingly being implemented at the government [1], funding agency [2, 3, 4], and journal [5, 6] level. These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8, 9, 10, 11]. However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by article age. For papers where the authors gave the status of their data, the odds of a data set being extant fell by 17% per year. In addition, the odds that we could find a working e-mail address for the first, last, or corresponding author fell by 7% per year. Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data

The scientific method: data is key



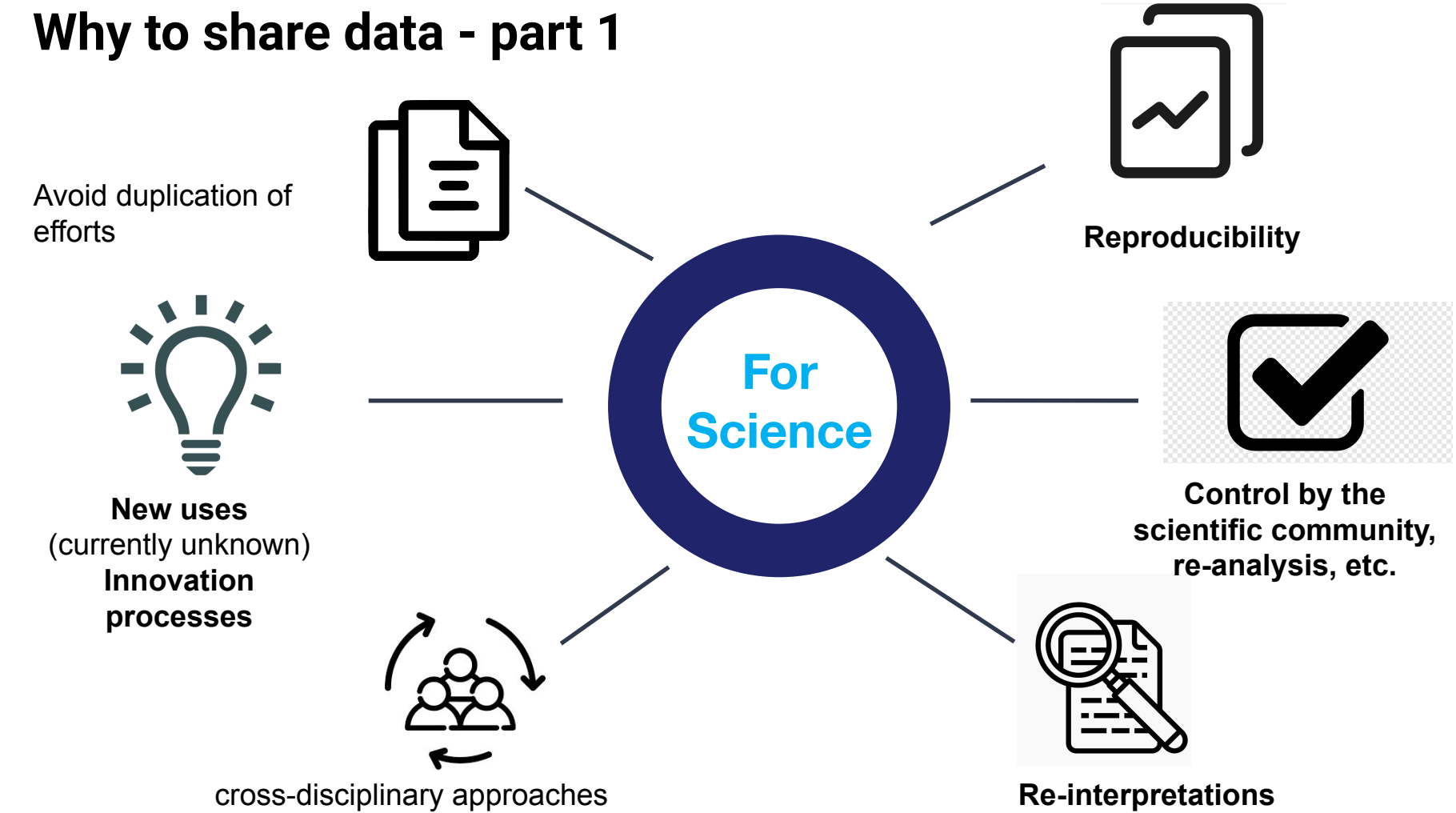
Portrait of Galileo Galilei by Justus Sustermans (Galleria degli Uffizi, Firenze), public domain

Why should you care?

- If you manage it, you probably will not **lose** it
- Organising your data will make your work more **efficient**
- Some data is **unique and not reproducible** (meteorology, observation from the field) so you should take care of it
- By correctly managing your data, you can improve **research integrity**
- By managing your data, you enable **validation and control**
- Someone else could use it in the future to **advance scientific progress**

Photo by Miguel Perales on Unsplash

Why to share data - part 1



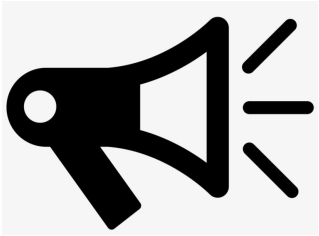
Why to share data - part 2

To comply with mandates

Publish a data paper

For your own research

More citations



More visibility and impact

Data reuse

stories & use cases

Legal Policy Webinars

Supporting researchers on the reuse of data: legal aspects to consider

Legal aspects on reusing data

Webinar aimed at PhD students / researchers

Recommendations for researchers on reusing data - legal aspects: what you need to know about GDPR and PSI Directive - common misperceptions and troubleshooting - practical examples.

How to advise researchers

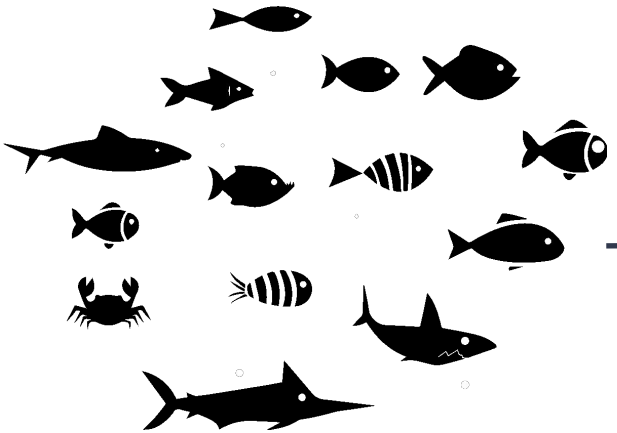
Webinar aimed at research support staff

What are relevant things to know when it comes to GDPR, new PSI directive and IPR - how to advise researchers - common misperceptions and troubleshooting - practical examples.

https://www.openaire.eu/data-reuse-use-cases

FishBase

Raw data released with a cc by-nc licence
(consortium 1)

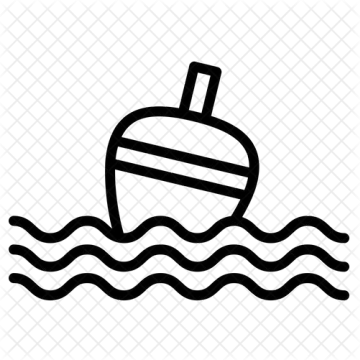


Institution two processes the data

A third consortium produces AquaMap

2527 bibliographical citation

Report on fishing stocks discussed at the European Parliament



Data from Marine floats (Argo)

Original raw data produced by Argo floats. Data are released with an open license (CC-0) on the institutional repository without any kind of processing.

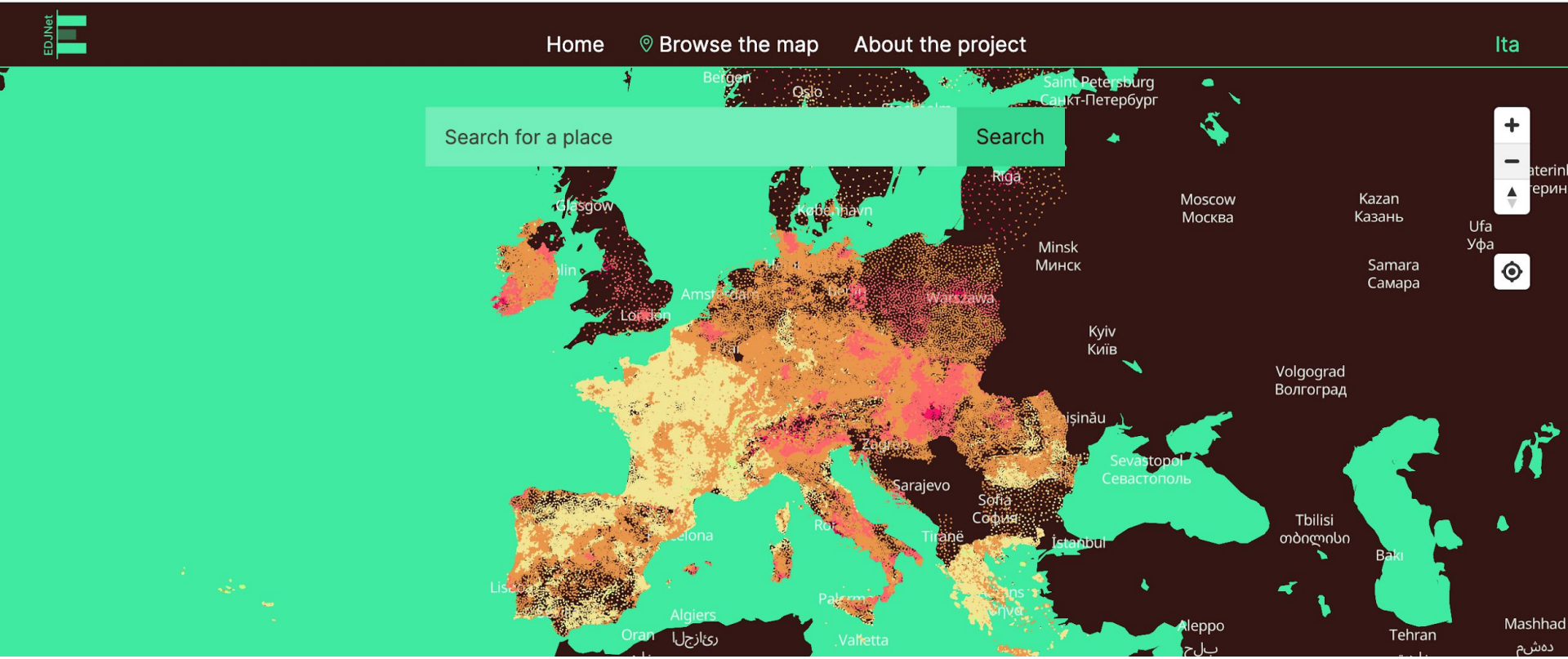
Standardization process

A standardization process was necessary and then a new dataset was produced.

Database: data released in CC-BY

The dataset is released with an open license (CC-BY) on the institutional repository and it is suggested to mention Institution 1 and Institution 2 in case of reuse.

Data reuse examples - a journalistic investigation



Based on data from the Copernicus project. https://climatechange.europeandatajournalism.eu/en/

About

The investigation was carried out by OBC Transeuropa for European Data Journalism Network (EDJNet) in 2020, as part of the network's long-standing interest in environmental issues and in the climate crisis. This article summarizes the main findings from a European perspective.

Methodology

The data analysis is based on data produced by UERRA regional reanalysis for Europe on single levels from 1961 to 2018, which was created by Copernicus and by the European Centre for Medium-Range Weather Forecasts (ECMWF). The data provides estimated temperatures values at two meters from the ground and it covers a cell grid, with cells 5,5×5,5 km large.

For reproducibility and visibility

https://www.data.cam.ac.uk/intro-data-champions/data-champions-cartoons

Hands on!

Exercise 1

Open Zenodo.org and find a record for software. Answer the following questions:

1. What is the DOI?
2. Is it linked to GitHub?
3. Which license applies?
4. Does it show versions?
5. How many views? Downloads?
6. Is it open access?

Time: 15 minutes

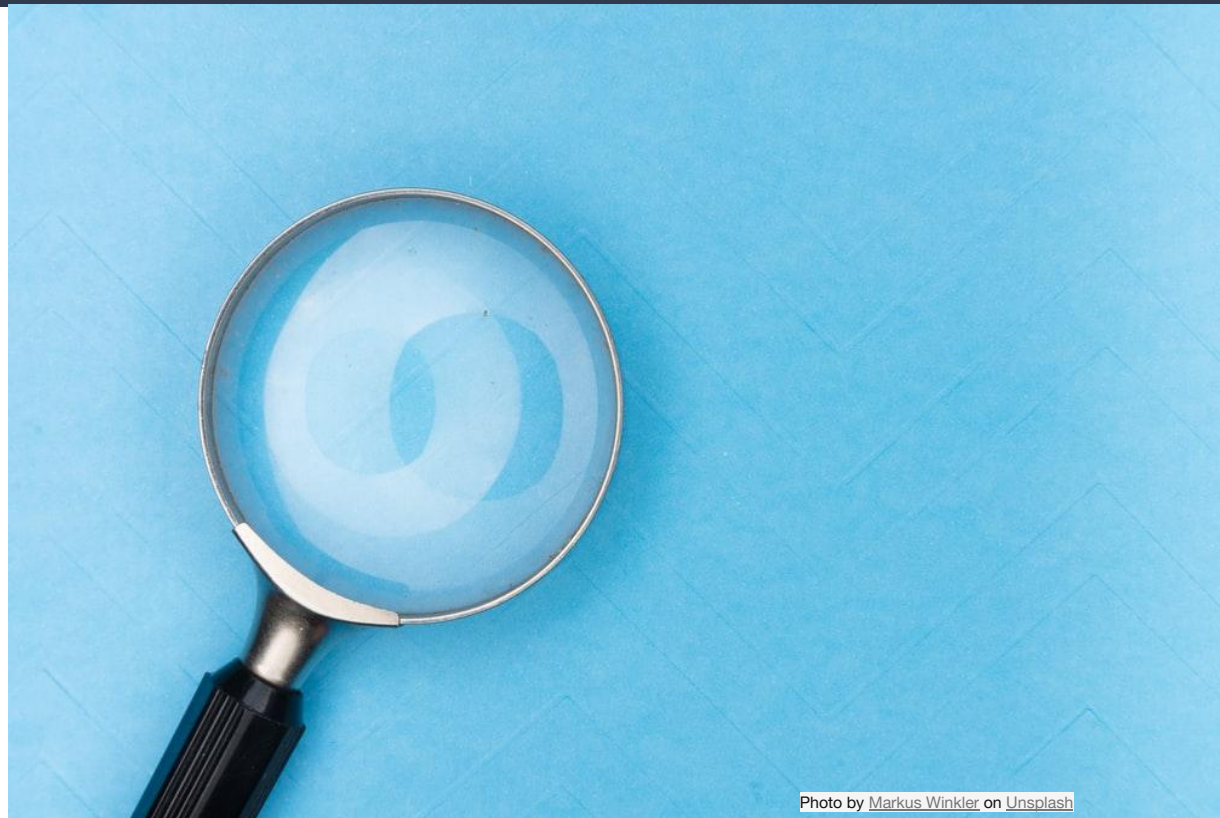
Credit: https://au-research.github.io/FAIR-data-101-training/findable/activities

Exercice 1: results

Did you find a record that met all the requirements?

Difficulties?
Doubts?
Questions?

Check Zenodo search guide:
https://help.zenodo.org/guides/search/




HAVE
A
BREAK
MACCHIATO
MACCHIATO

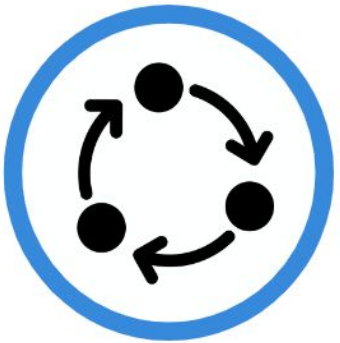
Exercice 2

Go to sandbox.zenodo.org and upload you file following FAIR principles

Findable

Accessible

Interoperable



Reusable

Time: 20 minutes

Exercice 2 results

Did you use the Orcid for the author(s)?

Did you notice that you can use communities to enhance findability?

Did you fill in the field “subject”?

Did you used keywords?

What field would you use if you are uploading a contribution to a conference?

Did you linked to other resources (ie. some sort of supplementary material or whatever)?

What if you want to add editors (or other contributors) to your uploand?

Exercice 3: assess your own record

Go to
https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/

Answer questions related to the principles underpinning Findable, Accessible, Interoperable and Reusable (FAIR)

What was your score?

Estimated time: 10 minutes

Exercice 4: take the quiz!

Try to answer to this quiz:

https://docs.google.com/forms/d/e/1FAIpQLSeHzqF8usfGGSoQPO5w46DNnTRXWIAvRkuWKz0lwF4U5uaQrg/viewform

The quiz reuses some of the following resources:
https://au-research.github.io/FAIR-data-101-training/

Thank you!

Ask questions and interact in the VRE:
https://services.d4science.org/group/phdunipi_os21-22

gina.pavone@isti.cnr.it

Consiglio Nazionale delle Ricerche

EOSC-Pillar

European Commission
Horizon 2020
European Union funding
for Research & Innovation