

# TAKE CARE OF DATA!

How Data Management and Open Science practices can improve your everyday life  
University of Padua, 2 December 2021

Gina Pavone, CNR-ISTI  0000-0003-0087-2151

# GINA PAVONE

- Research fellow at the Institute of Information Science and Technologies of the Italian National Research Council in Pisa, Italy.
- Research focus: Open Science and Open Access; Research Data Management
- OpenAIRE National Open Access Desk (NOAD) for Italy
- Coordinator of the editorial board of open-science.it
- My background: data journalism



# WHAT IS OPENAIRE

The European infrastructure for Open Science and Open Access.



## Mission:

Shift scholarly communication towards openness and transparency and facilitate innovative ways to communicate and monitor research.

## Services:

Provide interoperability services that connect research and enable researchers, content providers, funders and research administrators to easily adopt open science.

## Link research:

Link research outcomes (e.g., publications, data, software) to their creators (e.g., researchers, institutions, funders), enabling discoverability, transparency, reproducibility and quality-assurance of research.

<https://www.openaire.eu/>

# Outline

## Day 1

### 1 December

1

## Research

lifecycle and  
scholarly  
communication

2

## Open Science

What is it and  
why we need it

3

## RDM

Tips for everyday  
life with data

4

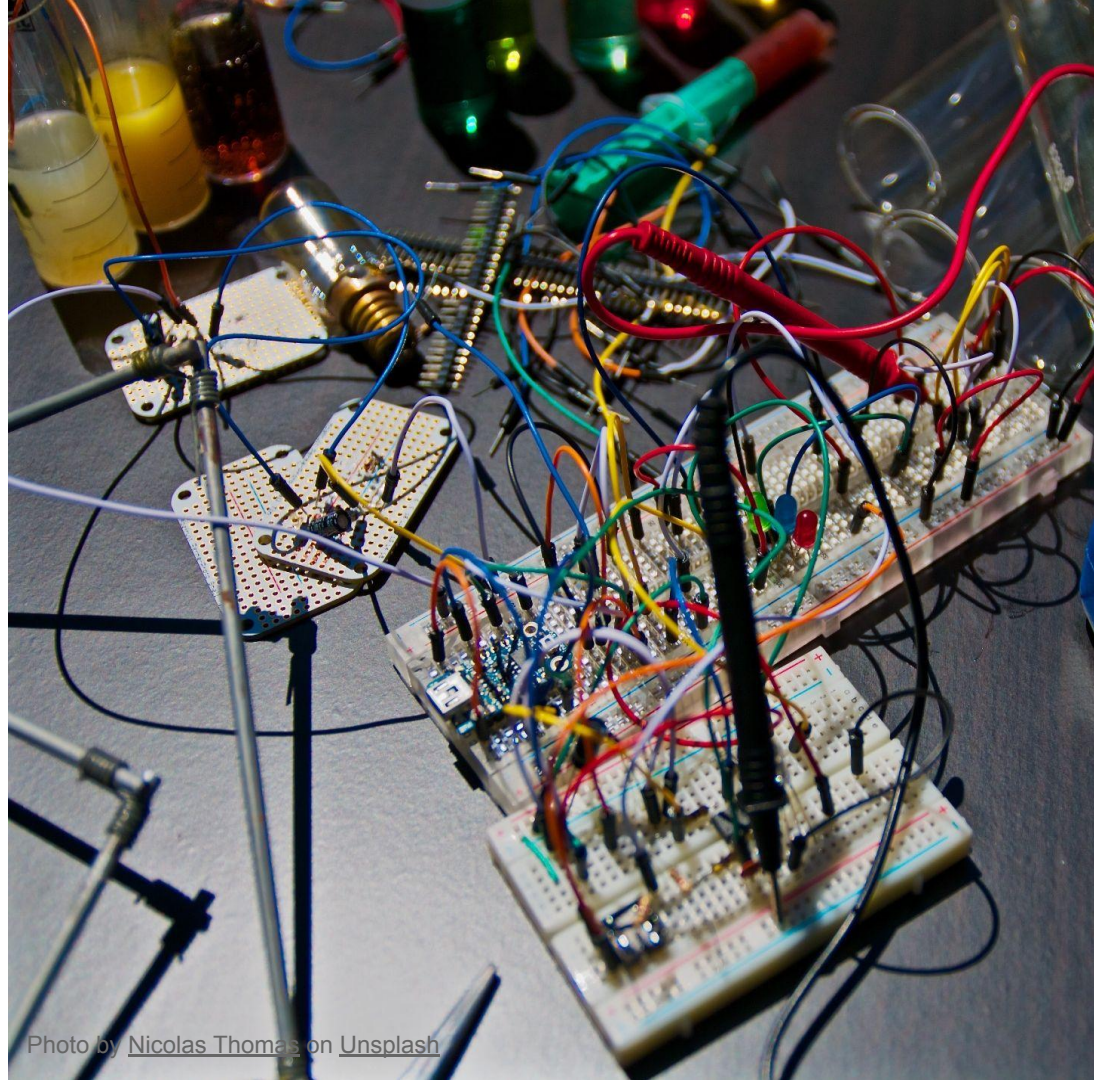
## FAIR

principles, to  
improve the  
value of data



## DO YOU HAVE A CLUE OF...

1. How scientific research works?
2. How scholarly communication works?
3. How many future PhDs?



# SCIENCE IS ALL AROUND IN SOCIETY

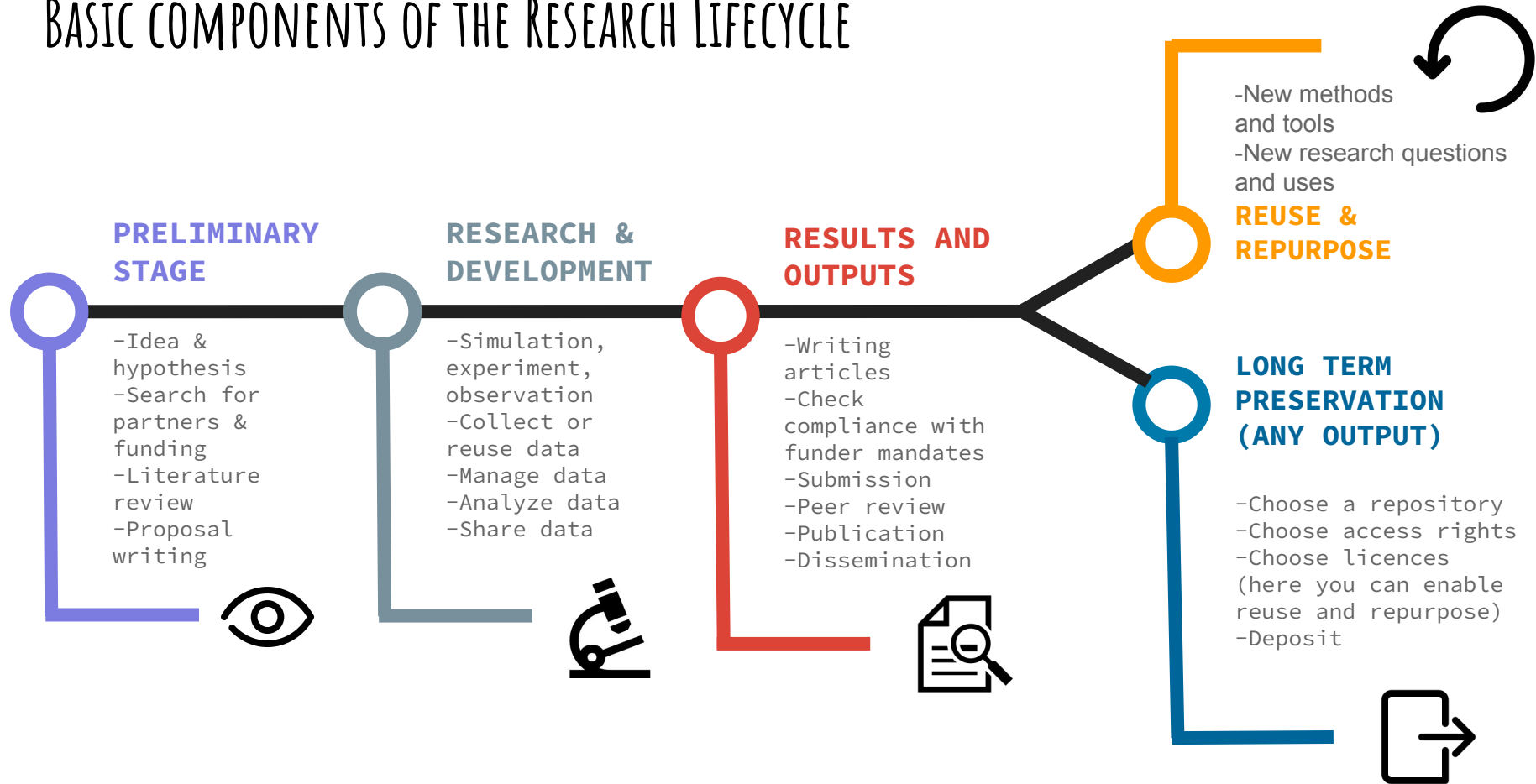
## What has CHEMISTRY ever done for you?

Chemistry is a significant contributor to the wealth, prosperity and health of society. Over the past 5,000 years, chemistry has fundamentally shaped our global civilisation.



Still curious? Find out more about chemistry at [science.org.au/curious/chemistry](https://www.science.org.au/curious/chemistry)  
 Download this infographic at [science.org.au/curious/chemistry-for-you](https://www.science.org.au/curious/chemistry-for-you)

# BASIC COMPONENTS OF THE RESEARCH LIFECYCLE









# publication

/ˌpʌblɪˈkeɪʃ(ə)n/

*noun*

the preparation and issuing of a book, journal, or piece of music for public sale.

"the publication of her first novel"

**Similar:**

issuing

announcement

publishing

printing

notification

reporting



- the action of making something generally known.

"the publication of April trade figures"

- a book or journal issued for public sale.

plural noun: **publications**

"scientific publications"

**Similar:**

book

volume

hardback

paperback

title

work

tome

opus



View PDF



Access through your institution

Purchase PDF

Search ScienceDirect



Outline

Highlights

Abstract

Keywords

- 1. Introduction
- 2. Background
- 3. Interview study and survey
- 4. Findings from the interview study and survey
- 5. Discussion
- 6. Conclusion

Declaration of Competing Interest

Acknowledgments

Appendix. Invitation letter to the survey

References

Show full outline



Journal of Systems and Software

Volume 182, December 2021, 111068



On researcher bias in Software Engineering experiments ☆

Simone Romano <sup>a</sup>, Davide Fucci <sup>b</sup>, Giuseppe Scanniello <sup>c</sup>, Maria Teresa Baldassarre <sup>a</sup>, Burak Turhan <sup>d, e</sup>, Natalia Juristo <sup>f</sup>

Show more

+ Add to Mendeley Share Cite

<https://doi.org/10.1016/j.jss.2021.111068>

Get rights and content

Highlights

Part of special issue:

SI: Recent Trends in Engineering Software-Intensive Systems

Edited by Antonio Martini, Manuel Wimmer

Download special issue

Recommended articles

Testing multiple linear regression systems with ...  
Journal of Systems and Software, Volume 182, 2021, Ar...

Purchase PDF

View details

Using metamorphic relations to verify ar...  
Journal of Systems and Software, Volume 182, 2021, Ar...

Purchase PDF

View details

Custom-tailored clone detection for IEC 61131-...  
Journal of Systems and Software, Volume 182, 2021, Ar...

Purchase PDF

View details

FEEDBACK

Full Text

Help

IF YOU WANT TO READ THIS PAPER YOU HAVE TO PAY. DIRECTLY OR INDIRECTLY

[nature](#) > [nature reviews physics](#) > [perspectives](#) > [article](#)

Perspective | [Published: 28 September 2021](#)

# Visualizing big science projects

[Katy Börner](#) , [Filipi Nascimento Silva](#) & [Staša Milojević](#)

[Nature Reviews Physics](#) (2021) | [Cite this article](#)

**39** Accesses | **8** Altmetric | [Metrics](#)

## Abstract

The number, size and complexity of ‘big science’ projects are growing – as are the

 [Access through your institution](#)

[Buy or subscribe](#)

**Sections**

[Figures](#)

[References](#)

[Abstract](#)

[Code availability](#)

[References](#)

AND IN AN INCREASINGLY INTERNATIONAL AND CROSS-DISCIPLINARY CONTEXT?

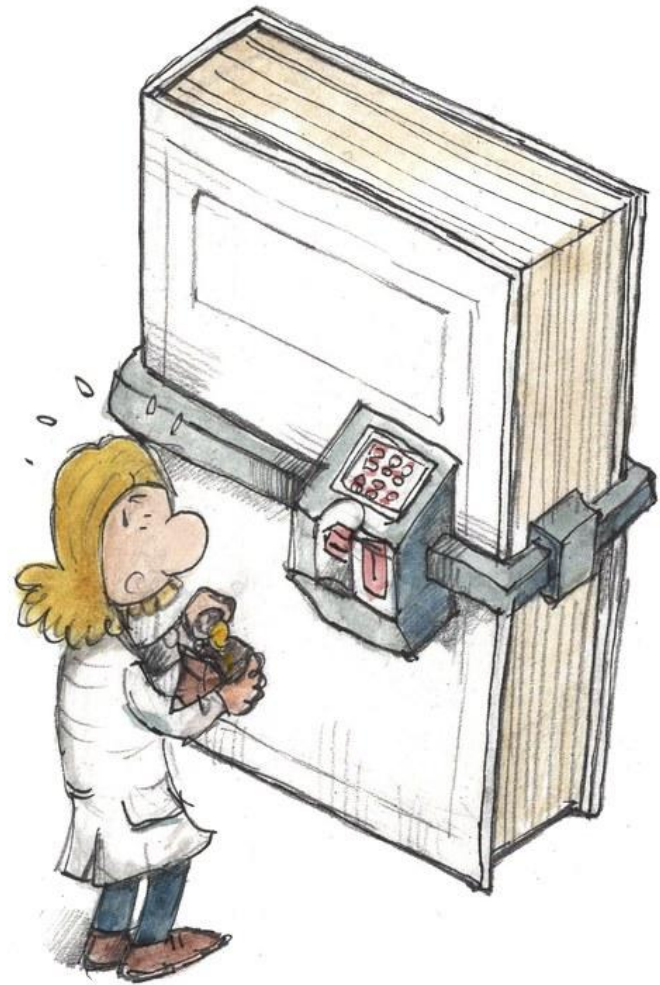
10 BILLION \$

The estimate of the global annual spending on academic journals throughout the world.



# KNOWLEDGE, BEHIND A PAYWALL !

One of the problems of the  
traditional scholarly  
communication system

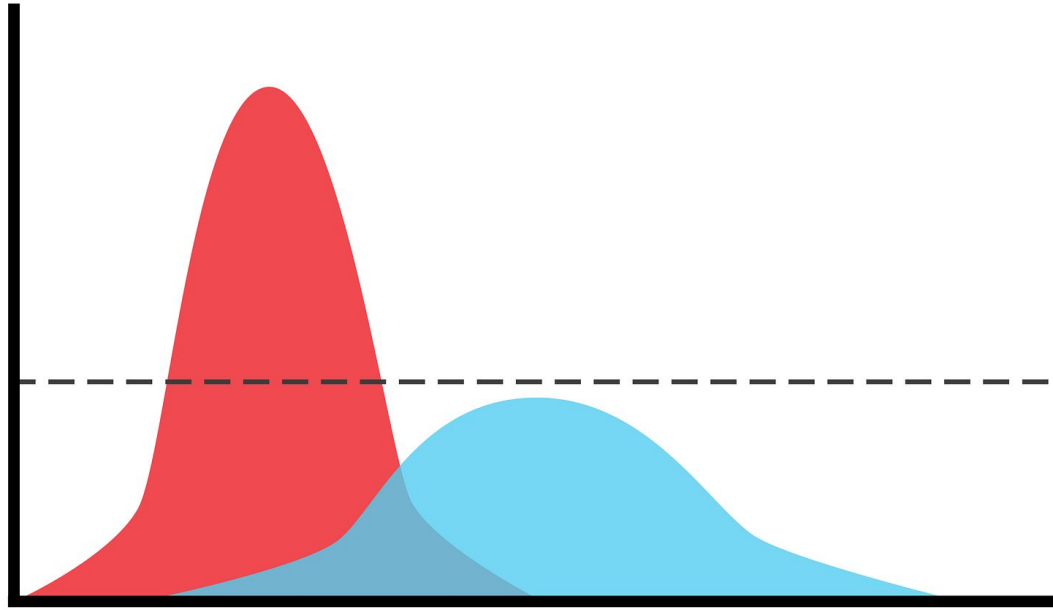


## WHERE DO RESOURCES COME FROM?

- Most part of the resources for science are publicly funded
- Science is in society and is intended for society



# THE CASE OF THE PANDEMIC: CURVES, DATA, MODELS... EVERYDAY LIFE ALSO FOR NON EXPERTS!



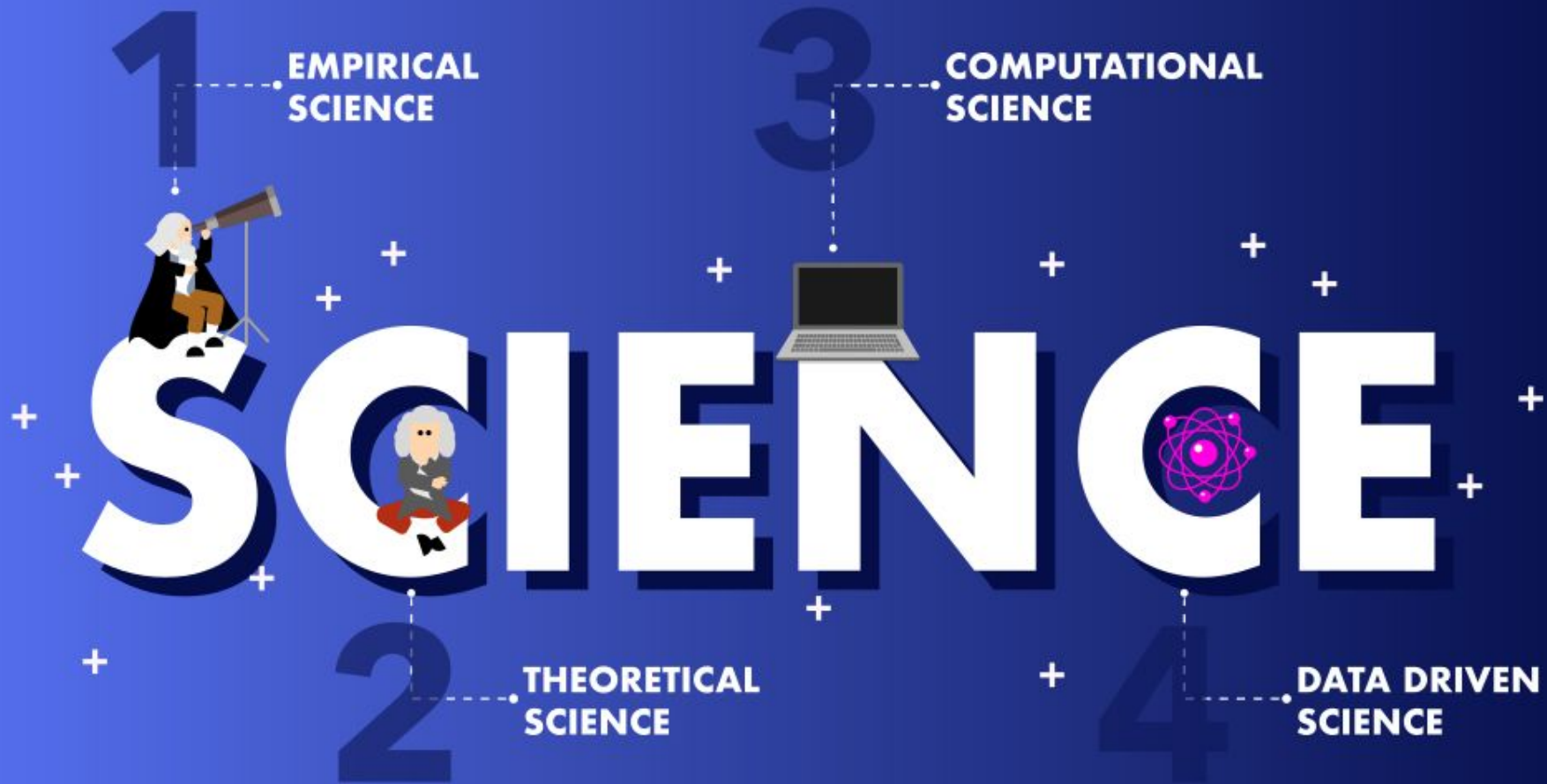
# THE PANDEMIC AND THE IMPORTANCE OF DATA

Public decision-making process and policies based on data.

The quality of data of crucial importance.

Good data management improve the quality of data and ultimately the quality of decisions affecting society





MAIN CONCEPTS OF  
SCHOLARLY  
COMMUNICATION

# PEER-REVIEW

One of the pillars of  
modern science

A formal “quality assurance” mechanism.

An author's scientific work undergo to the scrutiny of others who are experts in the same field (peers).

It is a check for soundness of scientific content, particularly from a methodological perspective.

# PREPRINTS

They can enable a more quick dissemination of research

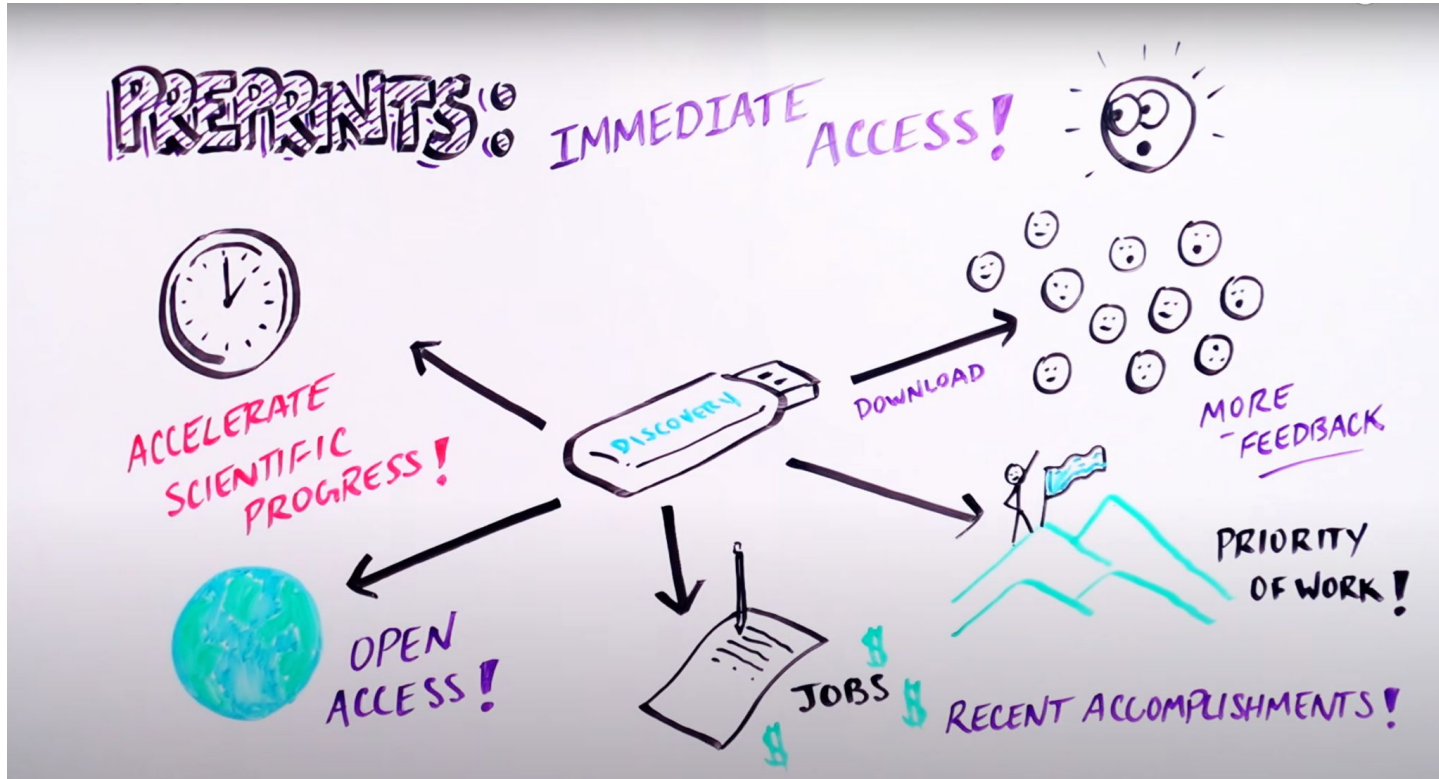
A preprint is the author's manuscript prior to peer review and formal publication.

Preprints can be submitted to a dedicated repository (like arXiv, bioRxiv, PeerJ, Zenodo)

---



# ADVANTAGES OF PREPRINTS



# POSTPRINTS

A postprint is peer reviewed scientific article

With editorial formatting:  
Version of record (VoR)

Without editorial  
formatting: Author  
Accepted Manuscript (AAM)

---

# DISSEMINATION

Researchers communicate in many ways their findings (scientific papers, conferences, repositories...)

Sharing research results with potential users - peers in the research field, industry, other commercial players, policymakers...and society at large!

---

WHY OPEN IS NOT THE DEFAULT MODE?





Scientific journals are  
subscription-based

## RESEARCH INSTITUTIONS

### PAY FOR:

- the work of researchers as **authors**
- the work of researchers as **reviewers**
- **access** to the results of the research work (they do not own anything!)





RESEARCHERS RELINQUISH THEIR COPYRIGHT TO PUBLISHERS/JOURNALS

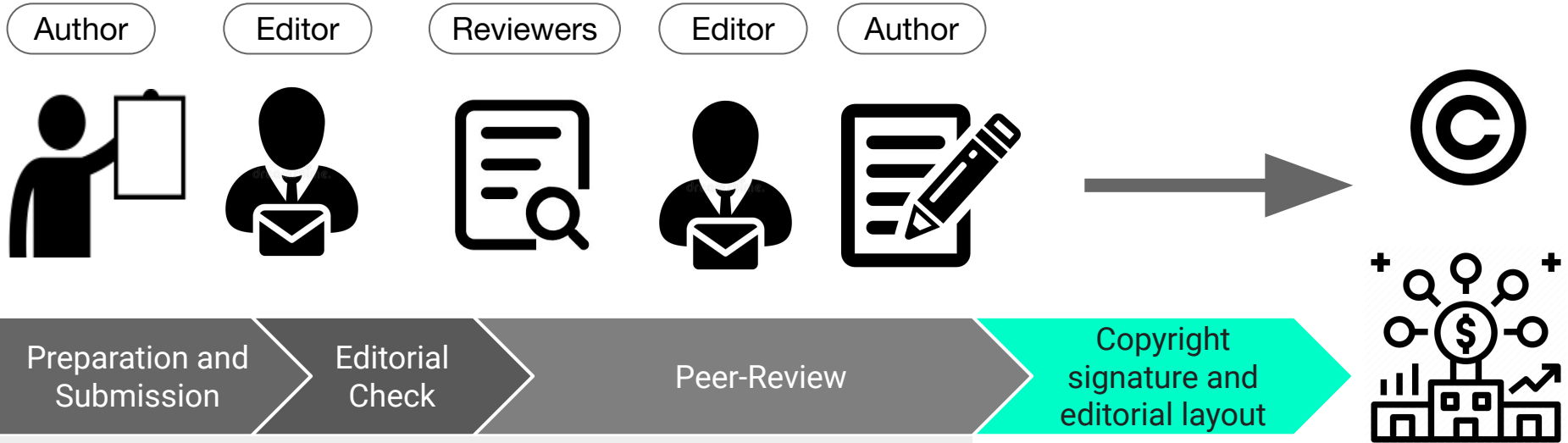


# AN INDUSTRY LIKE NO OTHER

In 2010, Elsevier's scientific publishing arm reported profits of £724m on just over £2bn in revenue. It was a **36% margin** – higher than Apple, Google, or Amazon posted that year.



# THE LIFE OF A SCIENTIFIC ARTICLE



Scientific community

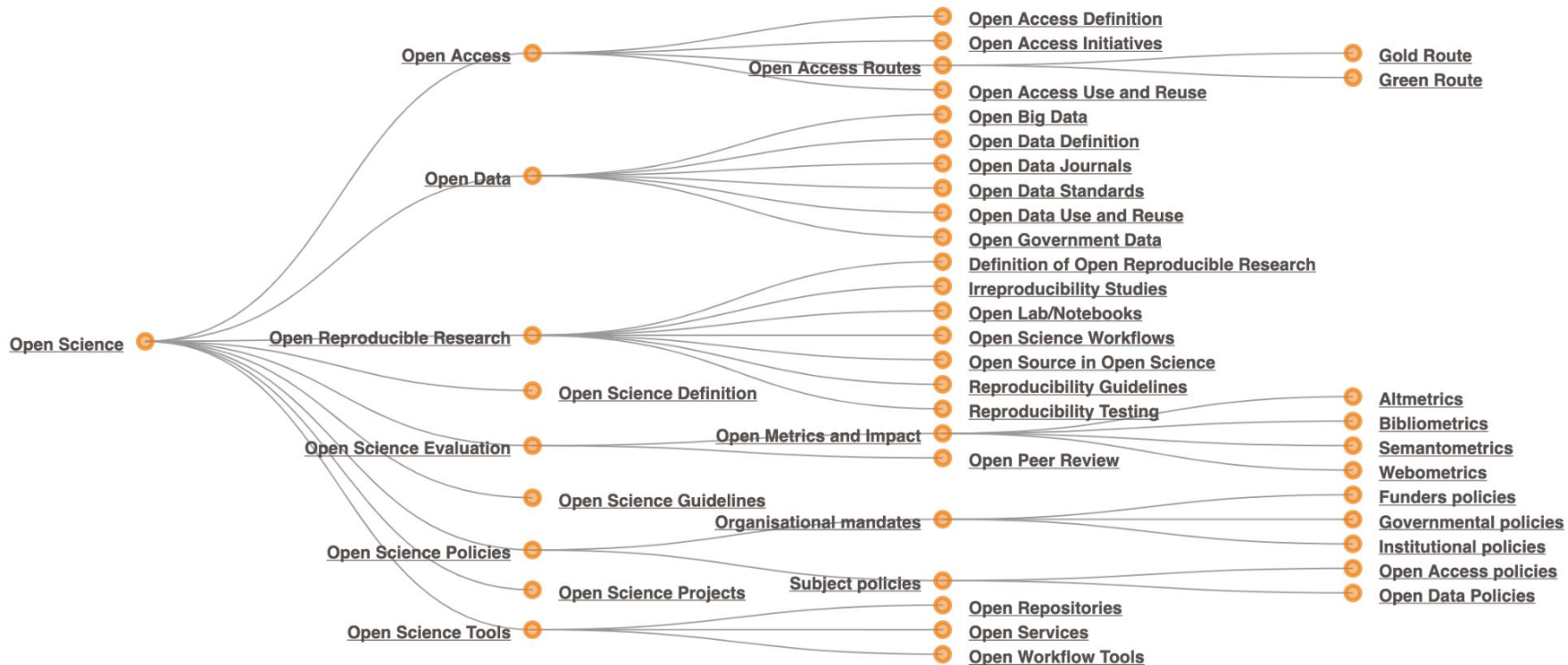
Commercial service



# WHAT IS OPEN SCIENCE?

Well, many many things...

# OPEN SCIENCE IS AN "UMBRELLA WORD"



# OPEN SCIENCE COMPONENTS

Open as much as possible  
**each step** of the research  
activity

[Unesco Open Science brochure](#)



# OPEN ACCESS

Research Outputs are  
available to anyone without  
costs or any other access  
barrier



# THE BENEFITS OF OPEN ACCESS

- Improving reach of research
- Helping to provide evidence for impact
- Improved reputation for researchers and their host institution through increased citations
- Improved quality of research through open, transparent and reproducible research practices



# RESEARCH INTEGRITY

- Transparency
- Collaboration
- Inclusion

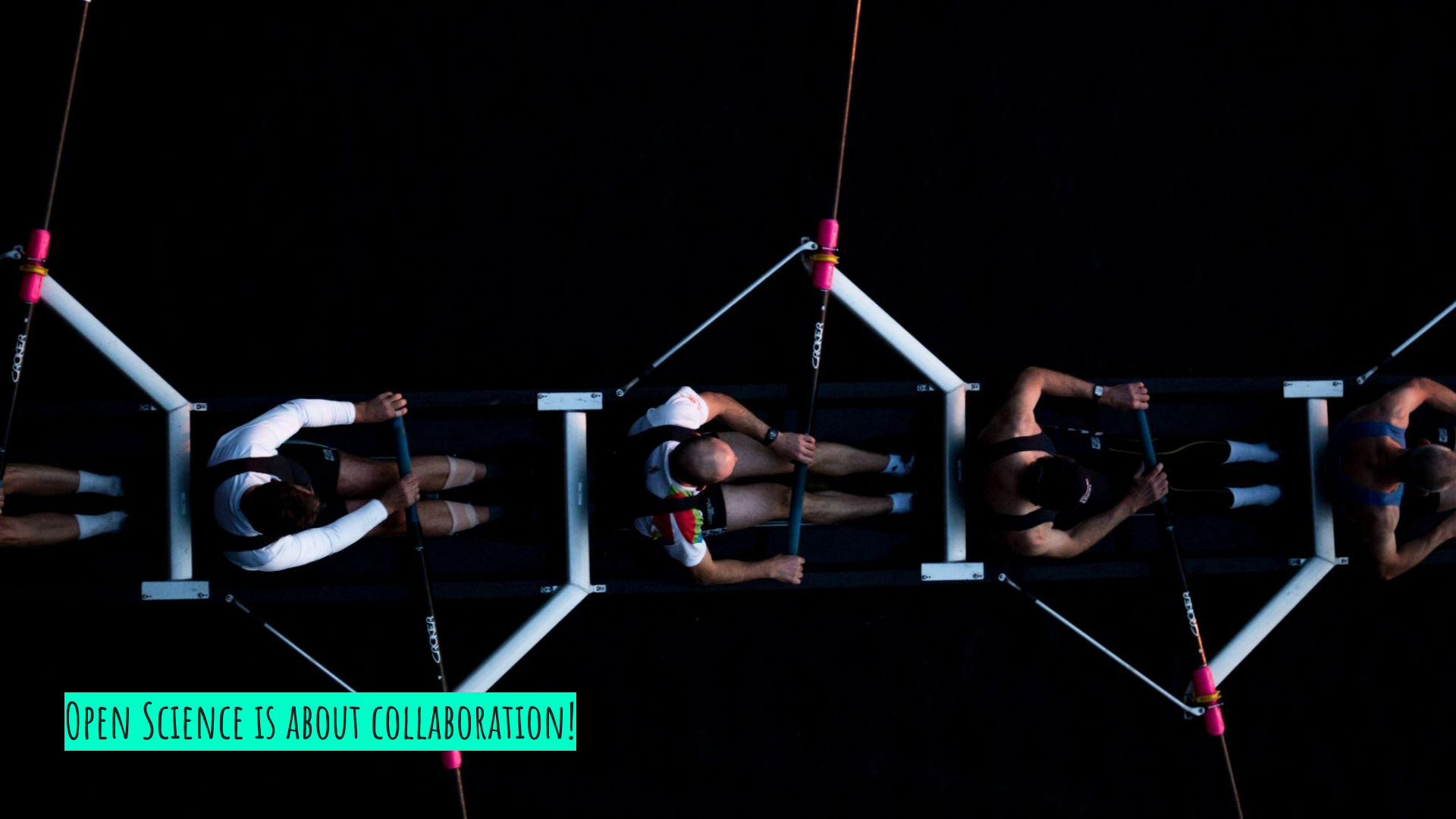




# OPENNESS TO DIVERSITY

- Recognize the diversity of knowledge systems and epistemologies
- Adhere to principles of non discrimination
- Availability of knowledge also for non wealthy countries





OPEN SCIENCE IS ABOUT COLLABORATION!



## THE IMPORTANCE OF NETWORKS

“

Open Science represents a novel approach to scientific development, based on cooperative work and information distribution through networks using advanced technologies and collaborative tools. Open Science seeks to facilitate knowledge acquisition through collaborative networks and encourage the generation of solutions based on openness and sharing.

”

# OPEN SCIENCE BENEFITS FOR SOCIETY AT LARGE

- More transparency
- Decisions based on better information (and decision-makers can more easily access and harness research results and methods)
- More opportunities for business: e.g. availability of scientific data for new applications.
- Increased awareness of the scientific methods and ways of working
- Improved scientific literacy: the general public can more easily access scientific results and methods.

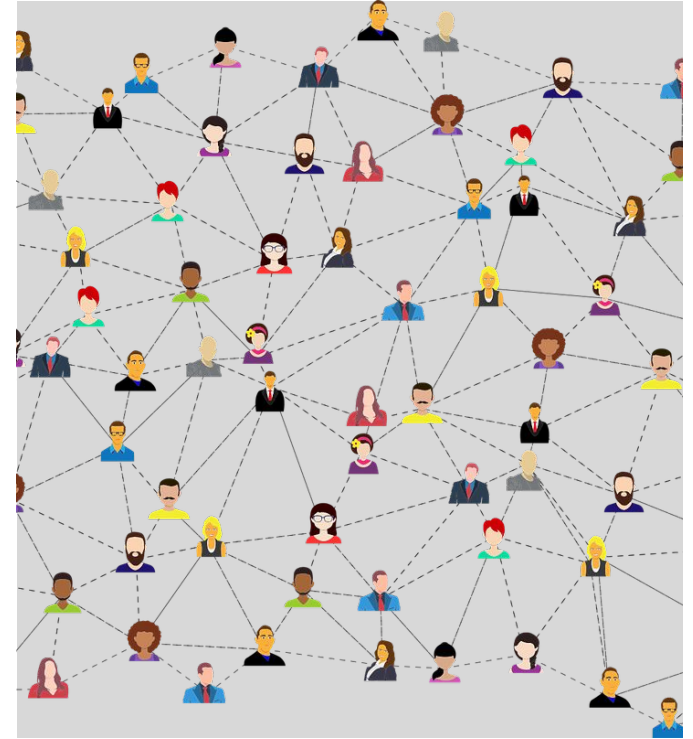


Image by Gordon Johnson from Pixabay

# CITIZEN SCIENCE

## Definition:

Projects and processes in which “citizen scientists, i.e. people who are not employed as scientists, can be involved in several stages of the scientific process, from collecting data to being part of the entire process from beginning to end”

<https://ec.europa.eu/jrc/en/science-update/science-citizen-science>

# CITIZEN SCIENCE

Citizens can participate in the scientific research process in different ways: as observers, as funders, in identifying images or analysing data, or providing data themselves.



# CITIZEN SCIENCE OBJECTIVE

Civil society participation  
in co-creating R&I content

The general public should  
be able to make  
significant contributions  
and be recognised as valid  
European science knowledge  
producers.

---

# CITIZEN SCIENCE PROJECTS

Currently there are citizen science projects and initiatives in almost every scientific field. Here just a few examples based in Italy

## Alien mosquito species



## Odour pollution



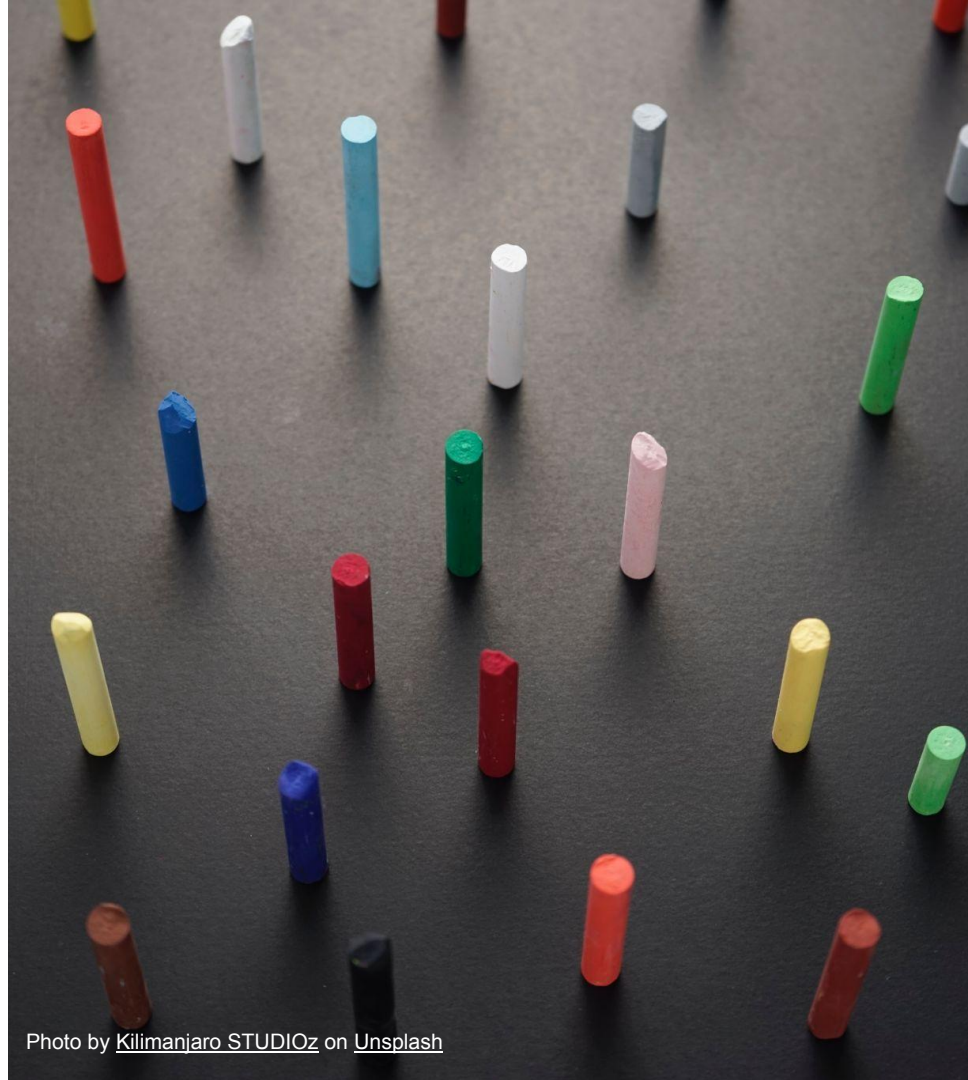
## Sensing for justice



# NOT ONLY PUBLICATIONS

Science is also:

- data
- software
- protocols
- negative results
- lab notes
- project deliverables
- and more...



26 BILLION €

ARE GOING LOST **EVERY YEAR** IN EUROPE FOR NOT  
MANAGING THE DATA PROPERLY



RESEARCH  
DATA  
MANAGEMENT

# WHAT HAPPENS IF...

Scientists 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

Subscribe



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<sup>6</sup>. 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.

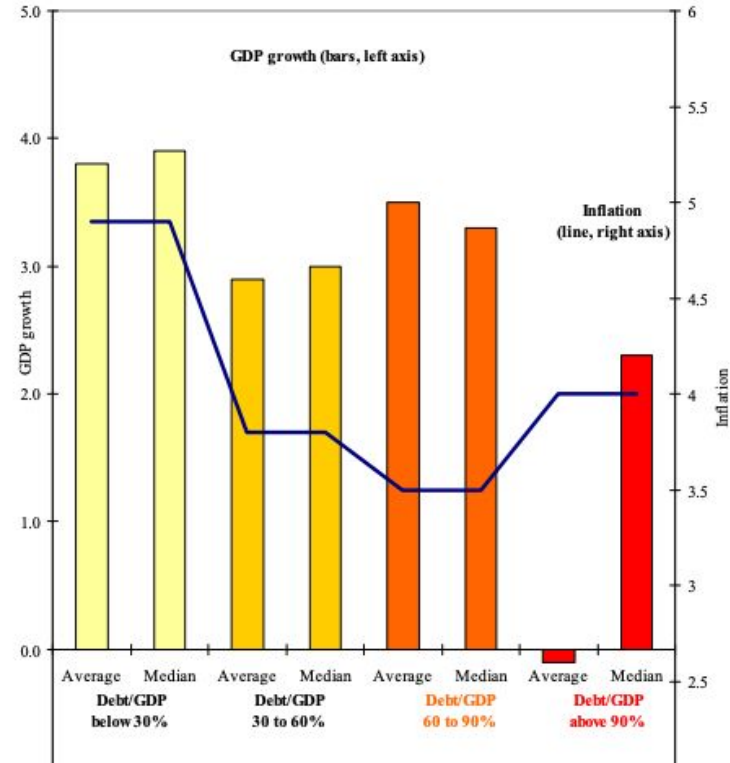
The five papers Myles warned about were quickly retracted, and the WHO revised its recommendation from 'strong' to 'conditional', meaning that clinicians have more freedom to make different choices for various patients. Schietroma says his calculations were assessed by an independent statistician and 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

# DATA MAY CONTAIN ERRORS

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.

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



# ERRORS AND MISCALCULATIONS

The screenshot shows a web browser displaying an article on 'THE CONVERSATION' website. The article title is 'The Reinhart-Rogoff error - or how not to Excel at economics'. The author is listed as 'Herndon, 2013'. The article text includes: 'Last week we learned a famous 2010 academic paper, relied on by political big-brothers to bolster arguments for austerity cuts, contained significant errors, and that those errors came down to misuse of an Excel spreadsheet. Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it? Harvard's Carmen Reinhart and Kenneth Rogoff are two of the most respected and influential academic economists active today. Or at least, they were. On April 16, doctoral student Thomas Herndon and professors Michael Ash and Robert Pollin, at the Political Economy Research Institute at the University of Massachusetts Amherst, re-audited the results of their analysis of over 2000 papers by Reinhart and Rogoff, papers that also provided much of the grist for the 2011 bestseller *This Time Is Different*. Reinhart and Rogoff's work showed average real economic growth slows (a 0.11 decline) when a country's debt rises to more than 90% of gross domestic product (GDP) - and this 90% figure was employed regardless of political arguments over high-profile austerity measures. During their analysis, Herndon, Ash and Pollin'.

## 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

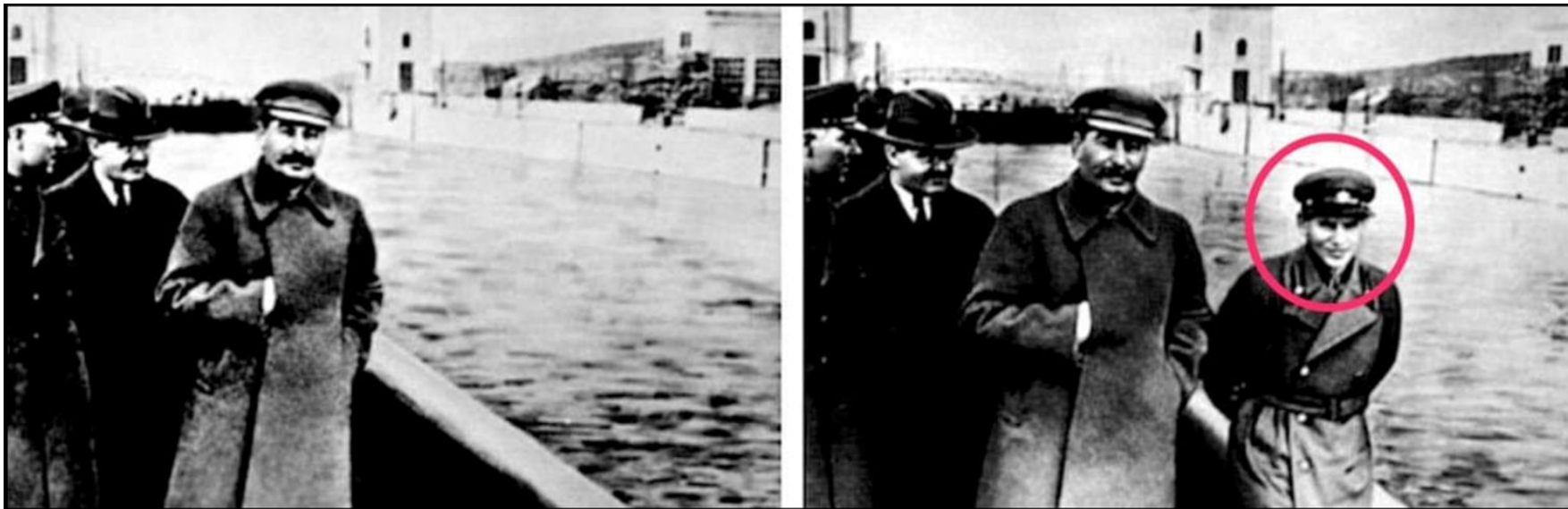
We replicate Reinhart and Rogoff (2010a and 2010b) and find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies 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](https://en.wikipedia.org/wiki/Nikolay_Yezhov)



# RESEARCH INTEGRITY: WE HAVE A PROBLEM

## REPORT

## Coping with Chaos: How Disordered Contexts Promote Stereotyping and Discrimination

Diederik A. Stapel<sup>1,\*</sup>, Siegwart Lindenberg<sup>1,2,\*</sup>

\* 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



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 RDM practices**

# REPRODUCIBILITY CRISIS

Availability of raw data  
underlying scientific  
publications falls by 17%  
per year

[https://www.cell.com/current-biology/fulltext/S0960-9822\(13\)01400-0](https://www.cell.com/current-biology/fulltext/S0960-9822(13)01400-0)

## The Availability of Research Data Declines Rapidly with Article Age

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

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



### Highlights

#### Summary

#### Reprints

#### 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

# WHAT IS DATA?

Data or it didn't happen!

Facts, observations or experiences on which an argument or theory is constructed or tested.

**Data are information!**  
(in a variety of forms and formats)

---

UCL Research Data Policy

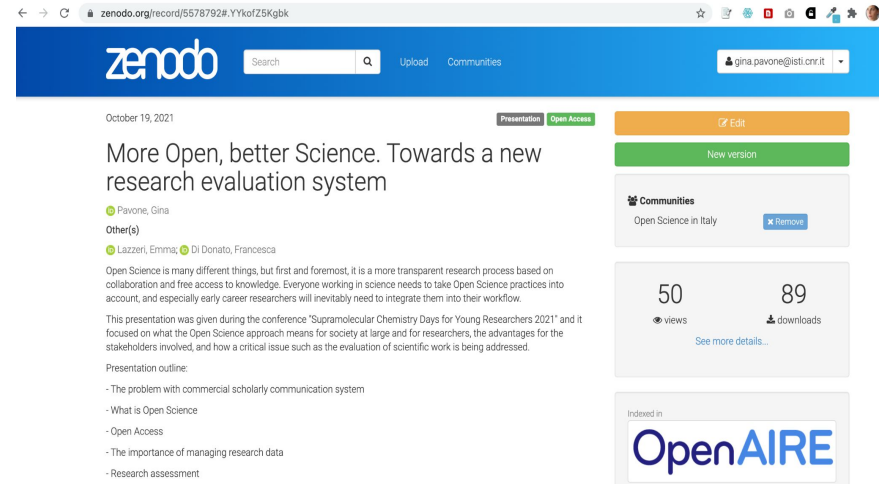
<https://www.ucl.ac.uk/library/research-support/research-data-management>



# REPOSITORY

An **open-access repository** or **open archive** is a digital platform that holds research output and provides free, immediate and permanent access to research results for anyone to use, download and distribute.

To facilitate **open access** such repositories must be **interoperable** according to the **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**.



The screenshot shows a Zenodo record page. The browser address bar displays the URL: zenodo.org/record/5578792#.YykoZ5Kgbk. The Zenodo logo is in the top left, with a search bar and 'Upload' and 'Communities' links. The user 'gina.pavone@isti.cnr.it' is logged in. The record is dated October 19, 2021, and is marked as a 'Presentation' and 'Open Access'. The title is 'More Open, better Science. Towards a new research evaluation system'. The author is listed as 'Pavone, Gina' with 'Other(s)' including 'Lazzeri, Emma; Di Donato, Francesca'. The abstract states: 'Open Science is many different things, but first and foremost, it is a more transparent research process based on collaboration and free access to knowledge. Everyone working in science needs to take Open Science practices into account, and especially early career researchers will inevitably need to integrate them into their workflow.' The record has 50 views and 89 downloads. It is part of the 'Open Science in Italy' community. The presentation outline includes: 'The problem with commercial scholarly communication system', 'What is Open Science', 'Open Access', 'The importance of managing research data', and 'Research assessment'. The record is indexed in OpenAIRE.

# KEY COMPONENTS

## Digital Object

In the context of scholarly communication: refers to any search result in its digital form, which can be uploaded into a repository (and possibly openly shared).

Examples: articles (pdf), datasets, software, images, videos, reports, posters or conference presentations, lectures (ppt), etc ...

## Metadata

Data about the DO

## Payload

The digital object (or digital objects) uploaded for deposit (and maybe shared). Also includes accompanying or description file (readme file, etc ...)

The screenshot shows a Zenodo repository page for a digital object. The page is titled "MOD01 - Research Data Management & Open Science: Introduction and Motivations - including EC policies and mandates" and is dated July 6, 2020. The page is part of a course held for a group of H2020 project participants. The page includes a search bar, a "New version" button, and a "Communities" section. The main content area shows the file name "INSI-Lazzeri\_MOD01\_20200706.pptx" with a size of 52.9 MB and a download button. Below the file information, there is a "Citations" section with a search bar and a "Show only" dropdown menu. The page also features a "Publications" section with a "DOI" field, a "Keywords" section, a "Grants" section, and a "Communities" section. The page is part of the OpenAIRE network. The page is also part of the Zenodo repository.



# METADATA

- Data describing data
- Very important for:
  - Access
  - Comprehension
  - Process
- Can be added manually or automatically
- There are discipline specific standards

innovation in metadata design, implementation & best practice

Join DCM!

## Dublin Core™ Metadata Initiative

Home Specifications News Community Learning About Contact

quick search...   

Home

The Dublin Core™ Metadata Initiative supports innovation in metadata design and best practices. DCMi is supported by its members and is a project of ASIS&T.

### Stewardship



For more than twenty years, the DCMi community has developed and curated [Dublin Core Specifications](#). More recently, DCMi has become recognised as a trusted steward of metadata vocabularies, concept schemes and other metadata artefacts, and has taken responsibility for other [community-created specifications](#). DCMi remains committed to this important work, and is actively developing more efficient and sustainable approaches to the stewardship of these standards, through the work of the [DCMi Usage Board](#).

### Learning



DCMi supports teachers and learners of modern metadata technologies and practices. An updated [Metadata Basics](#) page highlights current trends in descriptive metadata in the style of Dublin Core, which aims at interoperability through using globally shared vocabularies, constrained in application-specific profiles, based on principles of Linked Data. Interested learners can also explore a glossary page, a [Linked Data Competency Index](#) that enumerates relevant skills to be learned, a [guide for users of DCMi metadata terms](#), occasional [webinars](#) and tutorials at [DCMi annual conferences](#).

### Community



DCMi is defined by its [community](#) which is responsible for the innovative developments and evolving good practices which DCMi shares with the world. Much of DCMi's work is organised in [working and interest groups](#). DCMi's community is and has always been international, with active participants from around the world. The primary community event is the [DCMi Annual Conference](#). DCMi also organises regular [webinars](#), given by members of the community wishing to share their expertise with like-minded peers. Finally, DCMi [collaborates](#) with a number of other organisations.

### Development



DCMi has a long history of fostering and supporting technical development and innovation through the activities of its community, often in partnership with other organisations. Following on from the development of the ubiquitous [DCMi Metadata Terms](#), the community has in more recent years focussed on the concept of the [metadata application profile](#), developing supporting frameworks and conceptual models such as the [Singapore Frameworks](#). Most recently, the [Application Profiles Interest Group](#) has formed to address the next stage of development in this space.

Part of news image: © [Brett Womack](#) License: [CC-BY-NC-ND/3.0](#) Other photographs: © Paul Webb, License: [CC-BY/4.0](#)



DCMI 2020: Metadata Innovation  
Ottawa, Canada September 14th-17th, 2020

### News

#### DCMI 2020 Call for Proposals

Following on from the success of DCMi 2019 in Seoul (see Proceedings), we are pleased to announce the call for proposals in the DCMi 2020 International Conference on Metadata, Ottawa, Canada, 14-17 September 2020. We are grateful to Carleton University for offering to host us this year. This year's conference will mark the 25th anniversary of the original Dublin Core™ workshop. We will both reflect on two and a half decades of innovations while looking ahead to future developments. [read more...](#)

ISO 15836 Part 2 is published based on a revision of DCMi Metadata Terms

# PERSISTENT IDENTIFIERS

A persistent identifier (PI or PID) is a long-lasting reference to a document, file, web page, or other object.

The term persistent identifier is usually used in the context of **digital objects that are accessible over the Internet.**

Typically, such an identifier is not only persistent but actionable: you can plug it into a web browser and be taken to the identified source.

An example: DOI (Digital Object Identifier) aims to be **resolvable**, usually to some form of access to the information object to which the DOI refers.

This is achieved by **binding the DOI to metadata** about the object, such as a URL, **indicating where** the object can be found.

# RDM BENEFITS

## For researchers

- More visibility and citations
- Opportunity for collaboration
- Career recognition
- Helps to prevent errors and increases the quality of data analyses.
- Decrease of non-compliance risks (legal, ethical, institutional and funders' policies)

## For science

- Facilitates data finding and reuse
- Enables new research and new insights on the data
- Protection of valuable data
- Supports research integrity and reproducibility

## For society

- Efficient use of public resources
- Better quality research can benefit to better decision-making
- Opportunities for business
- Opportunities for citizen science
- Increased transparency and trust in science

RDM CAN BE USEFUL ALSO TO UNDERGRAD STUDENTS!



# IT IS A MATTER OF STRATEGY

- Managing (research) data is usually an integral part of the research process, **so you already do it**. You only have to **reflect on** and to **improve your strategy**.
- Most of the activities should be familiar:
  - **naming files** so you can find them quickly;
  - keeping track of different **versions**, and deleting those not needed;
  - **backing up** valuable data and outputs;
  - controlling who has **access** to your data.



# 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**





WHAT IS RDM AND HOW TO DO IT

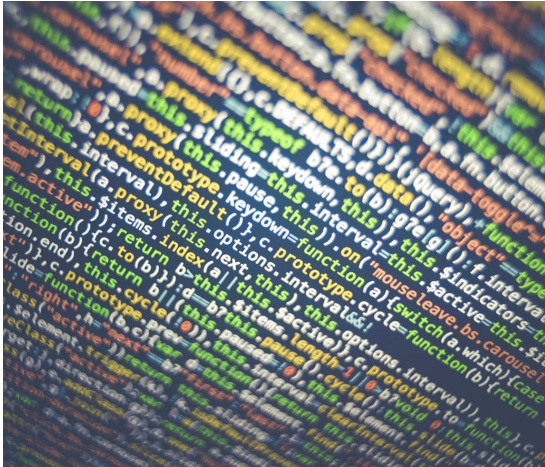
“Research data management (RDM) is the **careful handling and organization** of research data during the entire research cycle, with the aim of making the research process as **efficient** as possible and to facilitate cooperation with others. More specifically, RDM helps to **protect data**, it facilitates in sharing the data with others and it ensures that research data is findable, accessible and (re)usable”

Smits, D.A.B. and Teperek, M., 2020. Research Data Management for Master's Students: From Awareness to Action. *Data Science Journal*, 19(1), p.30. DOI: <http://doi.org/10.5334/dsj-2020-030>

# RESEARCH DATA MANAGEMENT

Actions and practices to ensure that research data are:

**Secure**



**Sustainable**



**(Re)usable**



Some of the following slides are inspired to the Ghent University guide on RDM: <https://www.ugent.be/en/research/datamanagement>

**Research Data Management** is simply the effective handling of information that is created in the course of research.

[How and why you should manage your research data: a guide for researchers](#)

[An introduction to engaging with research data management processes.](#)

Caroline Ingram, JISC Guides

# DATA ARE FIRST-CLASS RESEARCH OBJECTS

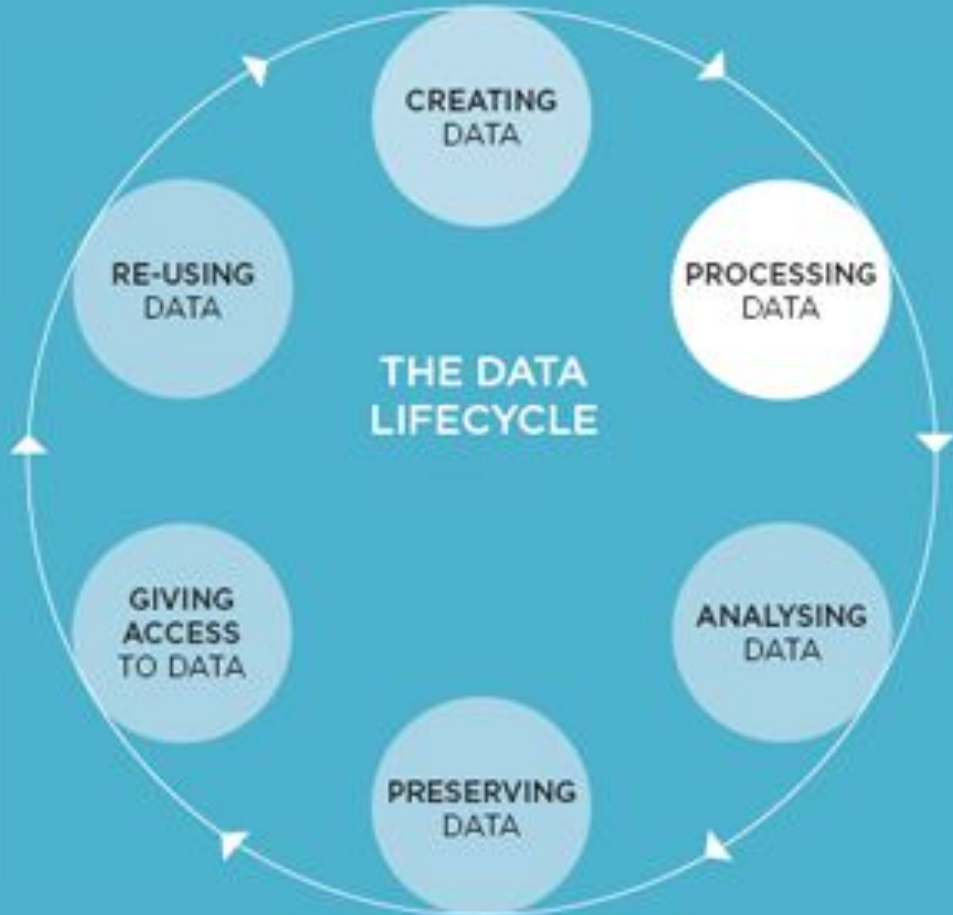
Check  
Validation  
Follow-ups  
New research questions  
Teaching  
Business applications  
...



PUBLICATIONS AND DATA

# RDM PRACTICES ENCOMPASS ACTIVITIES IN EVERY STAGE OF THE WORK WITH DATA

Before, during and after the research project. Choices made in one stage influence the next one.



# DATA COLLECTION

Consider what data will be collected, generated, and/or reused, and how you will organize them

Research data can be gathered through observation, manual or automatic measurements in the laboratory or in the field, with remote sensing techniques, by interviews, by modelling and simulation, etc. Data can also be stored in many formats.

This slide and some of the following are build reusing and manipulating parts of the Ghent University guide to RDM: <https://www.ugent.be/en/research/datamanagement> and/or <https://www.howtofair.dk/>



# DATA COLLECTION: FILE FORMATS

## File format:

- how information is stored within a digital file
- the format of a file is indicated by the 'extension' in the filename (e.g. .txt, .csv)

The choice of file formats to use depends on:

- Discipline-specific standards and customs
- Planned data analyses
- Software availability/cost
- Hardware used

## Risks

- Formats which can only be used within specific software makes the digital data vulnerable to obsolescence of the software
- you may have to keep some data files in multiple formats
- Beware to file converting!

## Best practices:

- Non-proprietary (not protected by trademark, patent or copyright).
  - Open, documented standard.
  - Common usage by research community
- Standard representation (ASCII, Unicode)
- include a readme.txt when using proprietary formats

# DATA COLLECTION: FILE NAMING

A file name is the principal identifier of a file.

Good file names should:

- Provide useful cues to content, status and version
- Uniquely identify a file
- Help to classify and sort files

File names can be constructed using the following elements:

- Project acronym
- Content description
- Date
- Location
- Creator name/initials
- Status information (i.e. draft or final) etc

# FILE NAMING: DECIDE WITH YOUR COLLEAGUES!



It can be useful if the consortium/department/group agrees on the following elements of a file name:

- **Vocabulary** - choose a standard vocabulary for file names, so that everyone uses a common language
- **Punctuation** - decide on conventions for if and when to use punctuation symbols, capitals, hyphens and spaces
- **Dates** - agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD
- **Order** - confirm which element should go first, so that files on the same theme are listed together and can therefore be found easily
- **Numbers** - specify the amount of digits that will be used in numbering so that files are listed numerically e.g. 01, 002, etc.

# FILE NAMING



# A GOOD EXAMPLE

[http://www.data.cam.ac.uk/files/gdl\\_tilSDocNaming\\_v1\\_20090612.pdf](http://www.data.cam.ac.uk/files/gdl_tilSDocNaming_v1_20090612.pdf)

## 3. Version

(upper case, max 4 chars, optional)

For documents that will continue in various versions use V followed by the version number. Use an underscore to indicate a decimal point if necessary.

Eg. PMF\_PRP\_ZenMonkeyProject\_V2\_20090607.docx

New versions should not be created for each iteration of the document, but rather at significant changes or when it has been reviewed or changed by another author.

Document naming for the TILS Division should follow this convention:

**GDL\_TILSDocNaming\_V1\_20090612.docx**

A prefix shows the document type

The document title describes the content

The version number

The date in the format yyyymmdd

Prefix	Meaning
AGD	Agenda
AGR	Agreement
GDL	Guideline
MEM	Memorandum
MIN	Minutes and Notes
PRE	Presentation
PRO	Procedure
PRP	Proposal
REP	Report
TEM	Template

## 2. Document title/ Description

(mixed case, max 30 chars, **no spaces**)

- Describes the purpose or “business” of the document. Acronyms, capitalisations, abbreviations can be used, keep in mind that descriptions should be **meaningful** to anyone reading the file name.
- In the case of project documentation use the **project name** or its usual abbreviation
- If possible Departmental Branch and/or Section should be integrated into this field to indicate origin / ownership of document.
- Use only alpha-numeric characters, plus the hyphen and underscore.
- **Do not use spaces.**

# DATA COLLECTION: FOLDER STRUCTURE

Information on a topic is located in one place

Are there established approaches in your team or department?

Name folders appropriately - i.e. name folders after the areas of work to which they relate

Structure folders hierarchically - limited number of folders for the broader topics, and then create more specific folders within these.

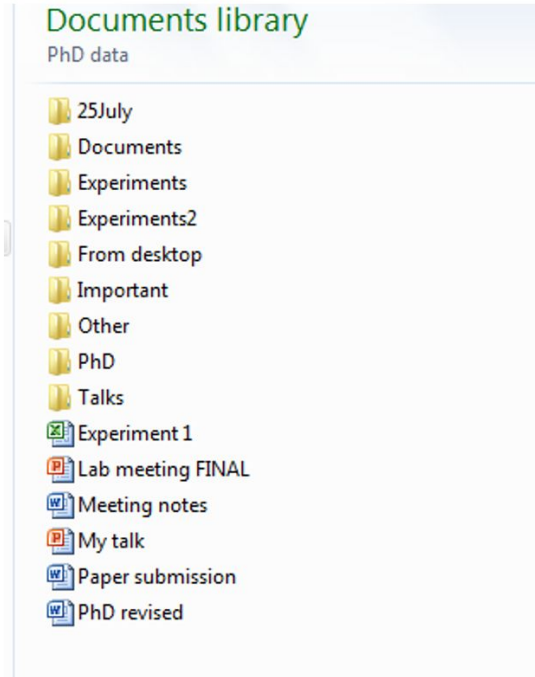
Consider separating ongoing and completed work.

Backup - ensure that your files, whether they are on your local drive, or on a network drive, are backed up.

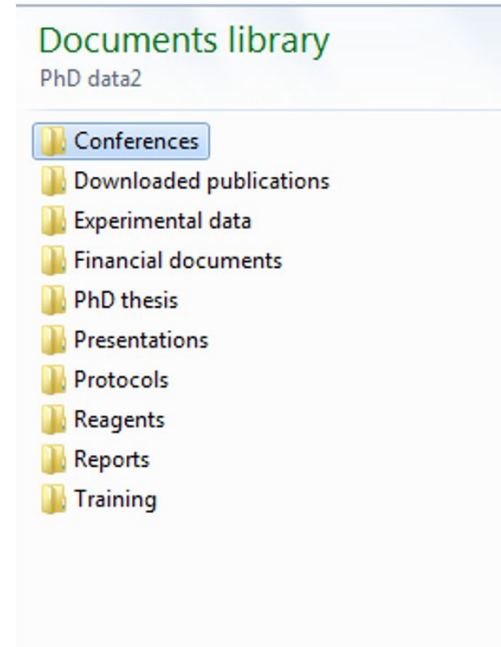
<https://www.data.cam.ac.uk/data-management-guide/organising-your-data>

# FOLDER STRUCTURE: WHAT IS YOUR STRATEGY?

## Example A

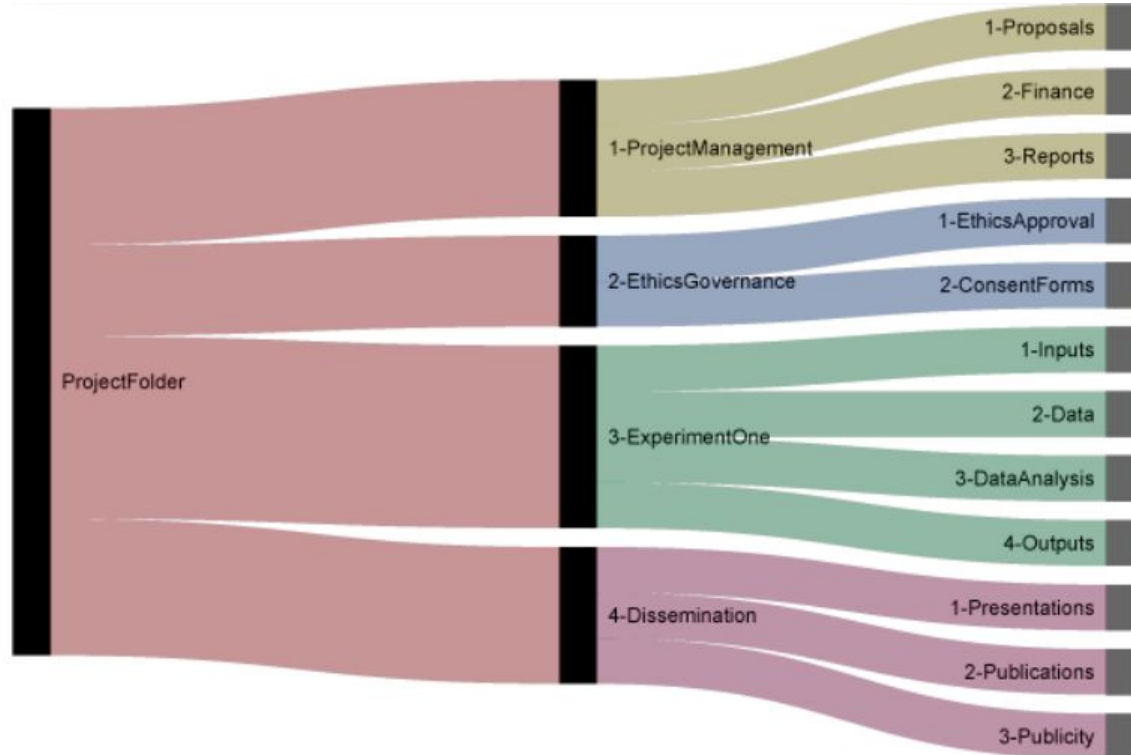


## Example B





# A GOOD EXAMPLE



# DOCUMENTATION

One of the basics of RDM. It enables you to understand/interpret data later

Study level documentation:

Contextual information (the background, aims, objectives, the hypotheses etc)

Procedural & methodological information

Data level documentation:

Information about datasets and/or individual data items

Information about variables, derived data, aggregated data etc

# SO MANY WAYS TO DESCRIBE YOUR DATA

How to create useful README files: <https://data.research.cornell.edu/content/readme>

```
Cornell AUTHOR_DATASET_ReadmeTemplate.txt

This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-----
GENERAL INFORMATION
-----

1. Title of Dataset

2. Author Information

Principal Investigator Contact Information
Name:
  Institution:
  Address:
  Email:
```

A readme file describes your data

Use a readme file for those data type that do not have a metadata standard available

README files template:

<https://cornell.app.box.com/v/ReadmeTemplate>

# LEGAL AND ETHICAL ASPECTS



Did you ask for an informed consensus to share the data and preserve them?

How will you protect personal data?

How about data licencing?

# DATA TO BE HANDLED WITH EVEN GREATER CARE:

- Personal data: any information about an identified or identifiable natural person (directly or indirectly)
- Personal sensitive data (i.e. revealing racial or ethnic origin, political views, religious or philosophical beliefs, membership of a trade union, genetic data, biometric data, data about health or someone's sexual behavior or sexual orientation)
- Data protected by IPR (Intellectual Property Rights) agreements
- Confidential data (i.e. commercial agreements)

This means that access to the data must be managed and restricted.

They still can be FAIR

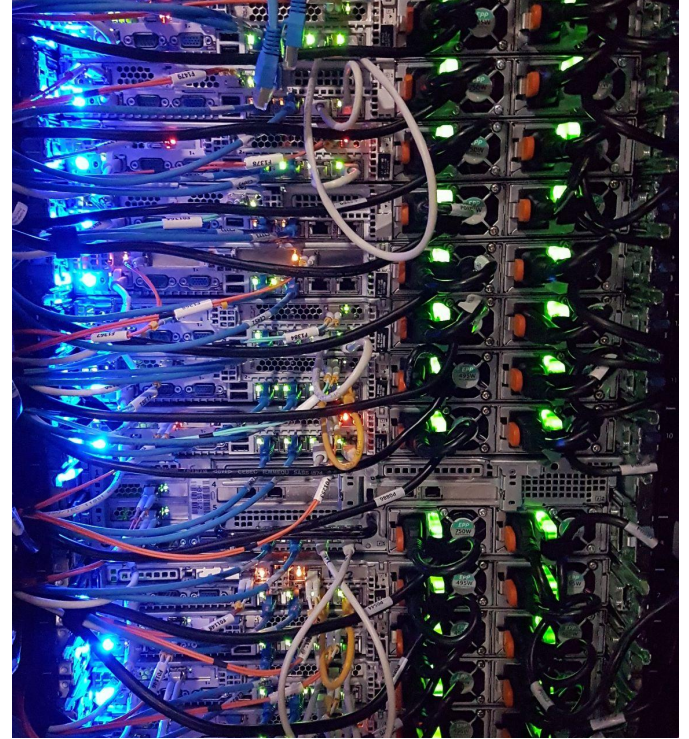
# DATA STORAGE AND BACKUP: DATA SHOULD BE SAFE!

Do you have enough space to store your data or should you include costs for additional services?  
Will the services be reliable/trusted?

How will you share your storage/backup with your collaborators?

Will you use cloud solutions?

Will you back your data up? How?





DO NOT LEAVE IT ALL TO GOOGLE

## Google services Terms of Use:

When you upload, submit, store, send or receive content to or through our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to

<https://policies.google.com/terms?hl=en>

# DATA SHARING

Make publicly available  
(outside project or research  
team)

Access: determining who you  
make your data available  
for, how you provide  
access, and under which  
conditions

Information on access  
conditions (if restricted -  
ie. email, form or  
whatever)

Use of persistent  
identifiers

Information on licences

---

# DATA PRESERVATION

Keeping data available and possibly usable in the longer term, beyond the end of the research project (deposit in a repository)

---

# SOFTWARE SHARING

GitHub



+



GitHub repositories can be deposited in Zenodo. This makes the repositories easier to reference in academic literature, creating persistent identifiers (DOIs).

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

# WHAT DOES FAIR MEAN



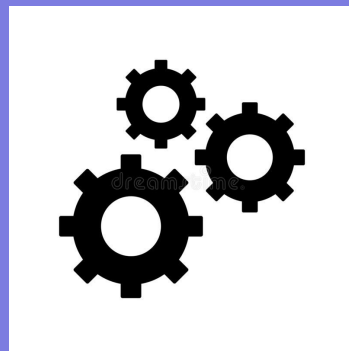
## Findable

The data is easy to find



## Accessible

It is clear who, when and how can access the data



## Interoperable

Data can be integrated with other data and/or they can be easily used and read by machines



## Reusable

Data can be reused by others in new research

# THE FAIR PRINCIPLES

www.nature.com/scientificdata

## SCIENTIFIC DATA

Amended Article

OPEN

SUBJECT CATEGORIES

• Research data

• Publication characteristics

### Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson et al.\*

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data. In addition to supporting its reuse by individuals, this Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

#### Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of long-term care of valuable digital assets, with the goal that they should be discovered and reused by downstream investigators, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—findability, accessibility, interoperability, and reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to the 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public); increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barbara.mon@ghis.it).

\*A full list of authors and their affiliations appears at the end of the paper.

SCIENTIFIC DATA | 8:166028 | DOI:10.1038/s41562-020-1518-9

- FAIR indicate a list of principles that can help you in making your data ready for Open Science
- They are **principles**, not standards!
- They were designed to enable optimal use of research data and methods
- A group of different experts designed the **FAIR principles** between 2014 and 2016
- They identified a set of 15 principles



GOOD PRACTICES  
TO MAKE  
YOUR DATA FAIR

# FAIRIFICATION BASICS

- **Documentation**

Gives the context to make your data understandable by others

- **Metadata**

Make your data easy to find

- **Data formats**

Make your data simple to combine to other data and machine readable.

- **Access to data**

It means to decide who will have access to your data and how

- **Persistent identifiers**

Persistent links to data that allows other to find and cite (give credit to) your data.

- **Licenses**

Are used to tell others how they can reuse your data.

- The first step in (re)using data is to find them.
- Metadata and data should be easy to find for both humans and computers.
- Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.



## Findable

F1. (meta) data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource



## Accessible

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

The ‘A’ in FAIR does not necessarily mean ‘open’ or ‘free’. Rather, it implies that one should provide the exact conditions under which the data are accessible. Hence, even heavily protected and private data can be FAIR.

- How do you give access to your data?

Through a Repository

- How do you choose the right repository?

Directory of Open access repositories:

[www.opendoar.org](http://www.opendoar.org)

Registry of Research Data Repository

<https://www.re3data.org/>

# Types of repositories

Who does curate/deposit in the repository?

## Thematic or disciplinary repositories

Designed for specific contents, curated by specific communities: ArXiv, bioarXiv, PMC...

[http://oad.simmons.edu/oadwiki/Disciplinary\\_repositories](http://oad.simmons.edu/oadwiki/Disciplinary_repositories)

## Institutional or national repositories

Maintained and curated by single institutions/countries. Typically only authors based in the specific institution/country can deposit, everyone can access

What are the repository contents?

## Literature Repositories

Reserved to text deposit (articles, reports, books, ...). Metadata reflect the repository contents.

<https://v2.sherpa.ac.uk/opensoar/>

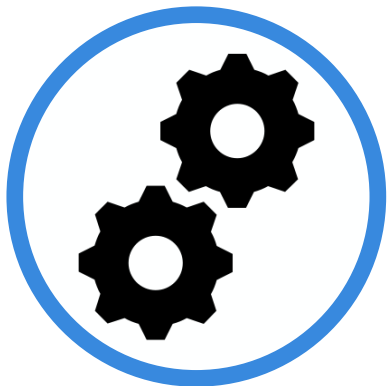
## Data repositories

Designed to deposit data. They often are disciplinary and have specific metadata to describe the type of data they preserve.

<https://www.re3data.org/>

## Catch-all repositories

All research products can be deposited (data, literature, presentations, poster, images, software, ...). Example: [Zenodo](#)



## **Interoperable**

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

Data usually needs to be integrated with other data.

In addition, data needs to interoperate with applications or workflows for analysis, storage, and processing.

**Use community standard or best practice!**

## Related/alternate identifiers

recommended 

Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

### Related identifiers

e.g. 10.1234/foobar.56789



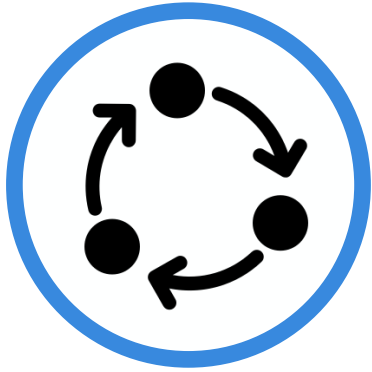
[+ Add another related identifier](#)

Optional. Resource type of the related identifier.

- ✓  cites this upload
- is cited by this upload
- is supplemented by this upload
- is a supplement to this upload
- is referenced by this upload
- references this upload
- is previous version of this upload
- is new version of this upload
- continues this upload
- is continued by this upload
- has this upload as part
- is part of this upload
- reviews this upload
- is reviewed this upload
- documents this upload
- is documented by this upload
- is compiled/created by this upload
- compiled/created this upload
- is the source this upload is derived from
- has this upload as its source
- is identical to this upload
- is an alternate identifier of this upload

- ✓  N/A
- Publication
  - Annotation collection
  - Book
  - Book section
  - Conference paper
  - Data management plan
  - Journal article
  - Other
  - Patent
  - Preprint
  - Project deliverable
  - Project milestone
  - Proposal
  - Report
  - Software documentation
  - Taxonomic treatment
  - Technical note
  - Thesis
  - Working paper
- Dataset
- Image





## Reusable

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

The ultimate goal of FAIR is to optimise the reuse of data.

To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings

## Licenses

Tell others how they can reuse your data!

## Provenance

Where is your data coming from?

# REFERENCES

DMP Tool: [https://dmptool.org/general\\_guidance#metadata-data-documentation](https://dmptool.org/general_guidance#metadata-data-documentation)

How to Guides DCC: <https://www.dcc.ac.uk/guidance/how-guides>

Five steps to decide what data keep:

<https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep>

Adapt your Data Management Plan. A list of Data Management Questions based on the

Expert Tour Guide on Data Management:

[https://www.cessda.eu/content/download/4302/48656/file/TTT\\_D0\\_DMPExpertGuide\\_v1.2.pdf](https://www.cessda.eu/content/download/4302/48656/file/TTT_D0_DMPExpertGuide_v1.2.pdf)

Practical Guide to the International Alignment of Research Data Management - Extended Edition, 2021, [10.5281/zenodo.4915861](https://doi.org/10.5281/zenodo.4915861)

OpenAIRE guide on cost in research data management,

<https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs>

Making Your Code Citable, <https://guides.github.com/activities/citable-code/>

Setting up an Organised Folder Structure for Research Projects,

[http://nikola.me/folder\\_structure.html](http://nikola.me/folder_structure.html)

TILS Document Naming Convention,

[http://www.data.cam.ac.uk/files/gdl\\_tilsdocnaming\\_v1\\_20090612.pdf](http://www.data.cam.ac.uk/files/gdl_tilsdocnaming_v1_20090612.pdf)

FAIRsharing, <https://fairsharing.org/standards/>

# REFERENCES

RDA, Metadata Standards Directory Working Group,  
<http://rd-alliance.github.io/metadata-directory>

Ghent University web guide on Research data management:  
<https://www.ugent.be/en/research/datamanagement>

Reading University web guide on RDM:  
<https://www.reading.ac.uk/research-services/research-data-management>

Francesca Di Donato, & Emma Lazzeri. (2021, October 18). Data Management. Zenodo.  
<https://doi.org/10.5281/zenodo.5593104>

Cambridge University, Data Management Guide: <https://www.data.cam.ac.uk/data-management-guide>

## REFERENCES

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

Martínez-Lavanchy, P.M., Hüser, F.J., Buss, M.C.H., Andersen, J.J., Begtrup, J.W. (2019). 'FAIR Principles'. In: Holmstrand, K.F., den Boer, S.P.A., Vlachos, E., Martínez-Lavanchy, P.M., Hansen, K.K. (Eds.), *Research Data Management* (eLearning course). doi: 10.11581 / dtu: 00000049

UCL Research Data Policy, <https://www.ucl.ac.uk/library/research-support/research-data-management>

# REFERENCES

C. Ingram, How and why you should manage your research data: a guide for researchers. An introduction to engaging with research data management processes, JISC Guides, <https://www.jisc.ac.uk/guides/how-and-why-you-should-manage-your-research-data>

B. Kramer, J. Bosman, Vienna Open Science workshop, 17-09-2020, [https://docs.google.com/presentation/d/1iw7zN\\_VS0mzTl81GnV0Itrjuvl2\\_1m0eXXlH3EFx4A8/edit#slide=id.p15](https://docs.google.com/presentation/d/1iw7zN_VS0mzTl81GnV0Itrjuvl2_1m0eXXlH3EFx4A8/edit#slide=id.p15)

K. den Heijer, CTG Data Management training day 1, 10.5281/zenodo.4048591

Fifty Shades of No, January 2015, [https://philarcher.org/diary/2015/50shadesofno/?fbclid=IwAR0gztaEKIjyn89n1azf4dFrYami\\_0hb-QSRKm\\_XpIRyzvLRRB929uCsckg](https://philarcher.org/diary/2015/50shadesofno/?fbclid=IwAR0gztaEKIjyn89n1azf4dFrYami_0hb-QSRKm_XpIRyzvLRRB929uCsckg)

Ghent University guide on Research Data Management: <https://www.ugent.be/en/research/datamanagement>

THANK YOU!

GINA.PAVONE@ISTI.CNR.IT

