

UNIVERSITY OF OSLO

Sharing and archiving

research data

Live Håndlykken Kvale & Agata Bochynska
University of Oslo Library
CC-BY-SA-4.0 2022

Research_data@uio.no



Agenda

- Repositories
 - Types of repositories
 - Finding repositories
 - Persistent identifiers
 - Archiving of code
 - Data paper
 - Selecting and preparing data
- Why archive research data?
 - Requirements
 - Data loss
 - Reasons for sharing
- How open?
 - Licenses
 - FAIR
 - Open science
 - Open formats
- Part two examples and menti



Academic twitter: What is the best way of making large phenotype data sets publicly available? Add to a repository, as supplemental data in a paper, or a different method?

Svar til 

One of my preferred remains in a repository (@figshare, @ZENODO_ORG, @datadryad), then add te doi of the dataset in the manuscript. I think it is easier to find the dataset that way (instead of going through the manuscript supplementals)

YES, a repository is the
best place to archive data

Domain-specific data repositories



General-purpose data repositories

The Zenodo logo consists of the word "zenodo" in a white, lowercase, sans-serif font, centered within a solid blue rectangular background.

zenodo



Open Science Framework

The Figshare logo features a circular pattern of multi-colored dots (red, green, blue, yellow) arranged in a ring, with the word "figshare" in a dark grey, lowercase, sans-serif font to its right.

figshare

National or institutional data repositories



NIRD RESEARCH DATA ARCHIVE

Finding repositories

Search

Browse ▾

Suggest

Resources ▾

Contact



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Search...

🔍 Search

Certification



<https://www.coretrustseal.org>
<https://www.fairsfair.eu/fair-certification>

Selecting an archive to deposit data

- Should the **data** be openly available?
- Should the **metadata** be openly available?
- Presence of **personal or confidential data** can affect choice of archive
- What can the relevant archives offer for **long term** perspectives?
- Does the archive offer **curation** – control of metadata and updating of formats?

Persistent identifier (PID)



<http://urn.nb.no/URN:NBN:no-5678>

ORCID

ROR

DOI – (digital object identifier) Commonly used for datasets and publications
URN – Persistent identifiers used in DUO for theses and self-archived publications.
ORCID – identifies the researcher
ROR - Identifies the University



Brea Manuel

@brea_manuel3

When it's your first paper and you're extreme. ❤️



Archiving of Code



scientific **data**

Data paper



Data in Brief
Open access

**Research Data Journal
for the Humanities
and Social Sciences**

<https://www.nature.com/sdata/>
<https://www.sciencedirect.com/journal/data-in-brief>
<https://brill.com/view/journals/rdj/rdj-overview.xml>

nature > scientific data

Reef Cover, a coral reef classification for global habitat mapping from remote sensing

Emma V. Kennedy, Chris M. Roelfsema ... Paul Tudman
Data Descriptor | 02 August 2021

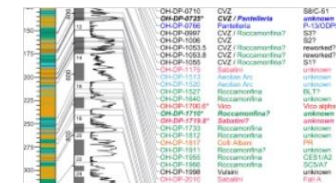


Featured

Data Descriptor
[Open Access](#)
02 Sept 2021

Lake Ohrid's tephrochronological dataset reveals 1.36 Ma of Mediterranean explosive volcanic activity

Niklas Leicher, Biagio Giaccio ... Bernd Wagner



17.03.2022

Data Descriptor
[Open Access](#)

OPERA tau neutrino charged current interactions



Selecting data for archiving

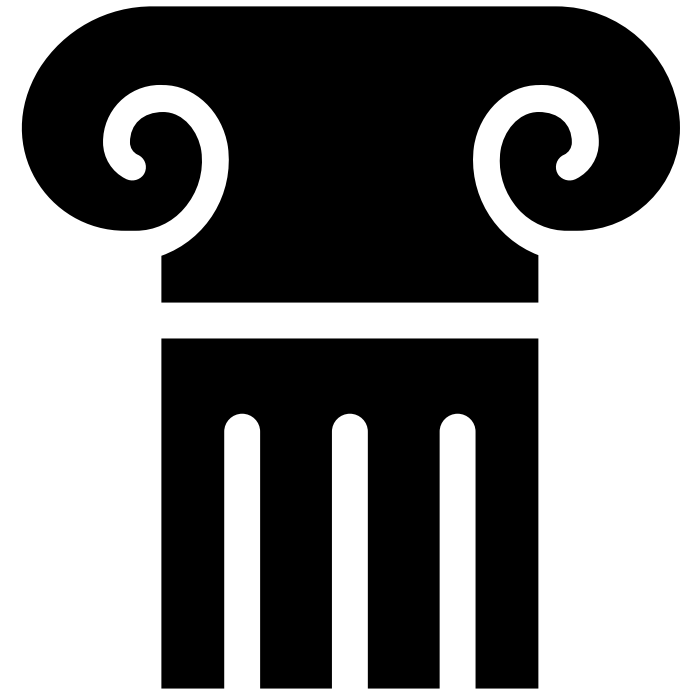
- Does your dataset have a potential for reuse?
- (Inter-)national or historical importance
- Data quality
- Uniqueness or originality
- Size, scale, cost
- Innovativeness

Preparing for archiving

- Consistent, meaningful, and compatible **file naming**
- Choose accessible, patent-free, and open **file formats**
- Make sure you have the necessary **documentation** (and metadata)
- **Reduce complexity** by grouping large groups of similar files in zip bundles to make upload and download easier
- Consider **size limitations** when choosing an archive (e.g. some archives have a limit of 10 to 50 GB per dataset)

Why is archiving important?

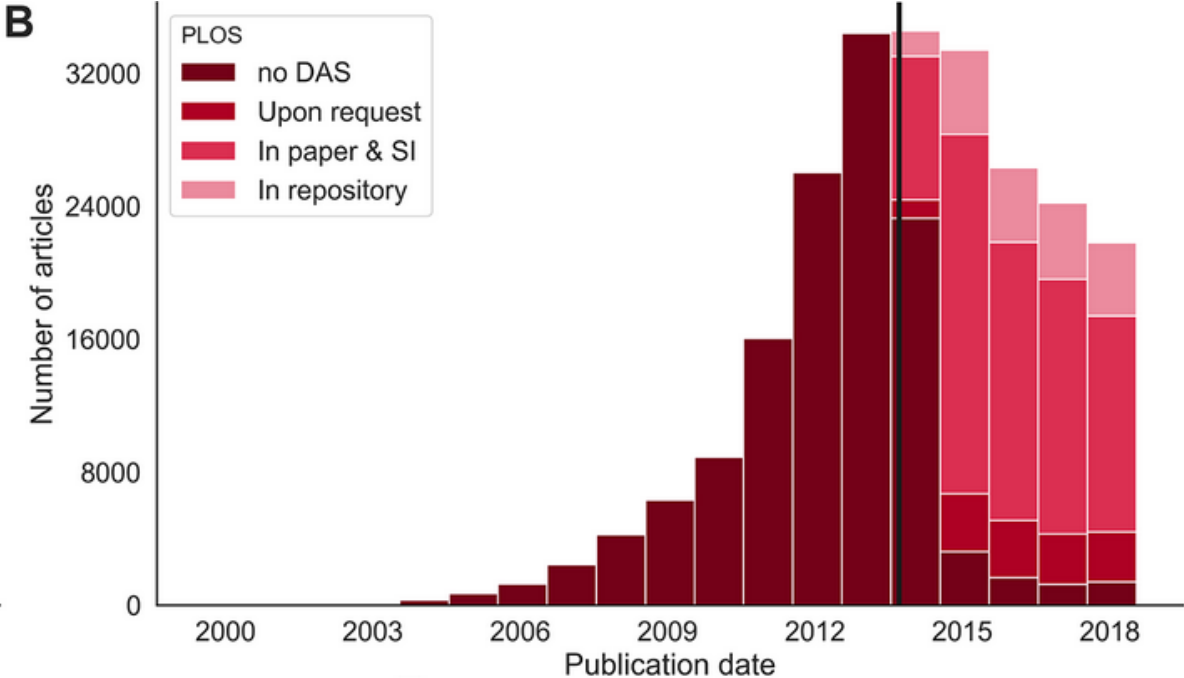
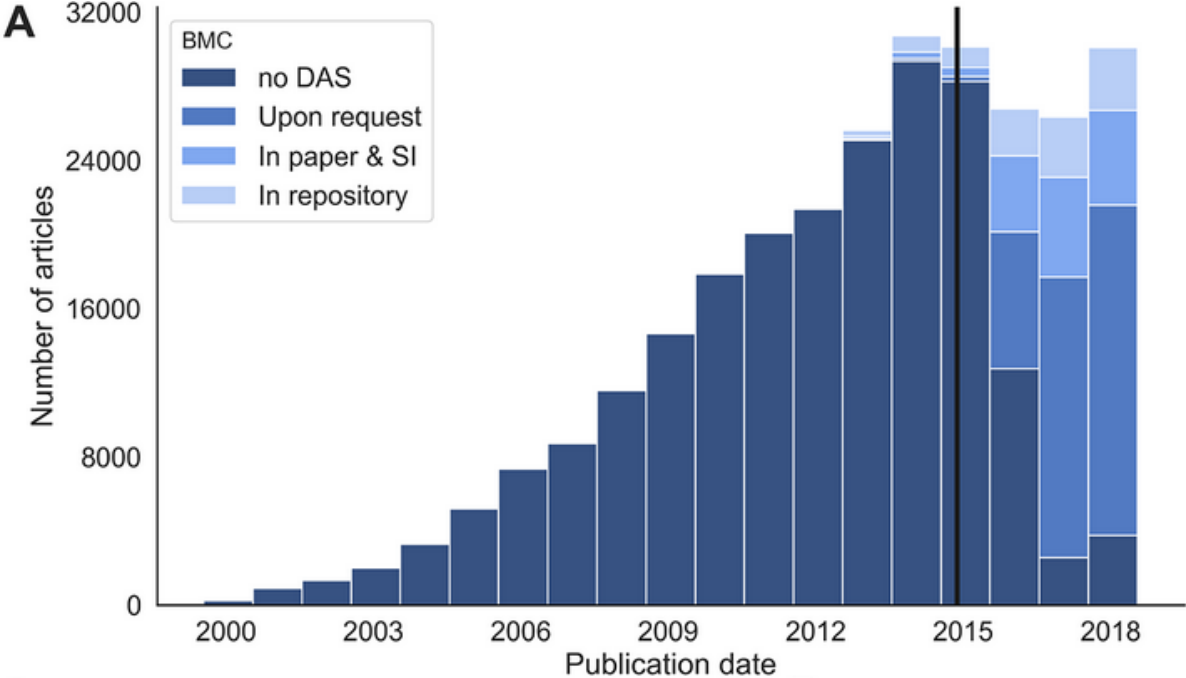
Requirements





“**Open Science** is becoming the modus operandi for carrying out research and innovation by **sharing knowledge, data and tools** as early as possible, in open collaboration with all relevant knowledge actors and society.”

Increasing number of deposited data



DAS: Data Availability Statement



MadScientist

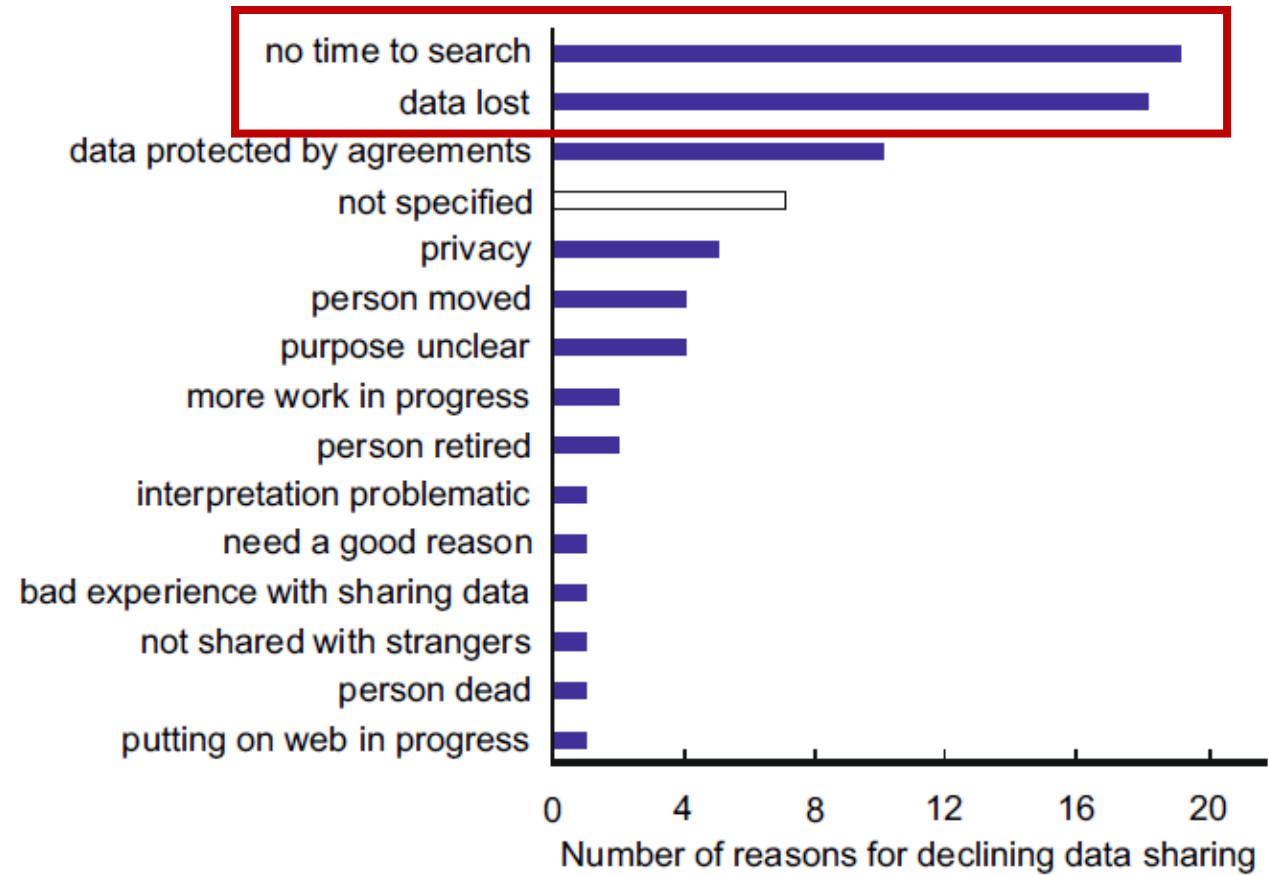
@MadS100tist



"Data will be available upon request"



Data availability statements don't work



The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,^{1,2,*} Arianne Y.K. Albert,³ Rose L. Andrew,¹ Florence Débarre,^{1,4} Dan G. Bock,¹ Michelle T. Franklin,^{1,5} Kimberly J. Gilbert,¹ Jean-Sébastien Moore,^{1,6} Sébastien Renaut,¹ and Diana J. Rennison¹

¹Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

²Molecular Ecology Editorial Office, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada

³Women's Health Research Institute, 4500 Oak Street, Vancouver, BC V6H 3N1, Canada

⁴Centre for Ecology & Conservation Biosciences, University of Exeter, Cornwall Campus, Tremough, Penryn TR10 9EZ, UK

⁵Institute for Sustainable Horticulture, Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, BC V3W 2M8, Canada

⁶Department of Biology, Université Laval, 1030 Avenue de la Médecine, Laval, QC G1V 0A6, Canada

Summary

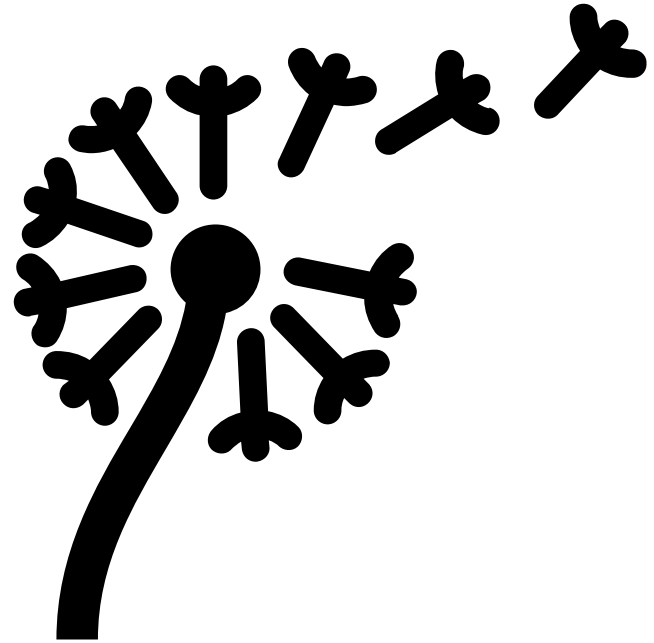
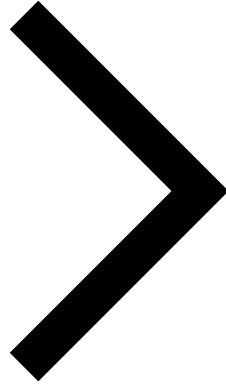
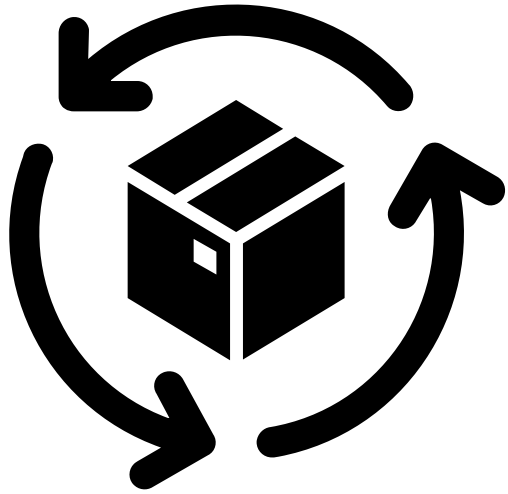
Policies ensuring that research data are available on public archives are increasingly being implemented at the government [1], funding agency [2–4], and journal [5, 6] level. These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8–11]. 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

sets (23%) were confirmed as extant. [Table 1](#) provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability that at least one e-mail appeared to work (i.e., did not generate an error message), (2) the conditional probability of a response given that at least one e-mail appeared to work, (3) the conditional probability of getting a response that indicated the status of the data (data lost, data exist but unwilling to share, or data shared) given that a response was received, and, finally, (4) the conditional probability that the data were extant (either “shared” or “exists but unwilling to share”) given that an informative response was received.

There was a negative relationship between the age of the paper and the probability of finding at least one apparently working e-mail either in the paper or by searching online (odds ratio [OR] = 0.93 [0.90–0.96, 95% confidence interval (CI)], $p < 0.00001$). The odds ratio suggests that for every year since publication, the odds of finding at least one apparently working e-mail decreased by 7% ([Figure 1A](#)). Since we searched for e-mails in both the paper and online, four factors contribute to the probability of finding a working e-mail: (1) the number of e-mails in the paper and (2) the chance that any of those worked and (3) the number of e-mails we could find by searching online and (4) the chance that any of those worked. The total number of e-mail addresses we found in the paper decreased with age (Poisson regression coefficient = -0.07 , SE = 0.01, $p < 0.0001$) from an average of 1.17 in 2011 to 0.42 in 1991 ([Figure 2A](#)), and there was a slight positive effect of article age on the number of e-mails we found online (Poisson regression coefficient = 0.015, SE = 0.007, $p < 0.05$; [Figure 2C](#)). Moreover, the chance that an e-mail found in the

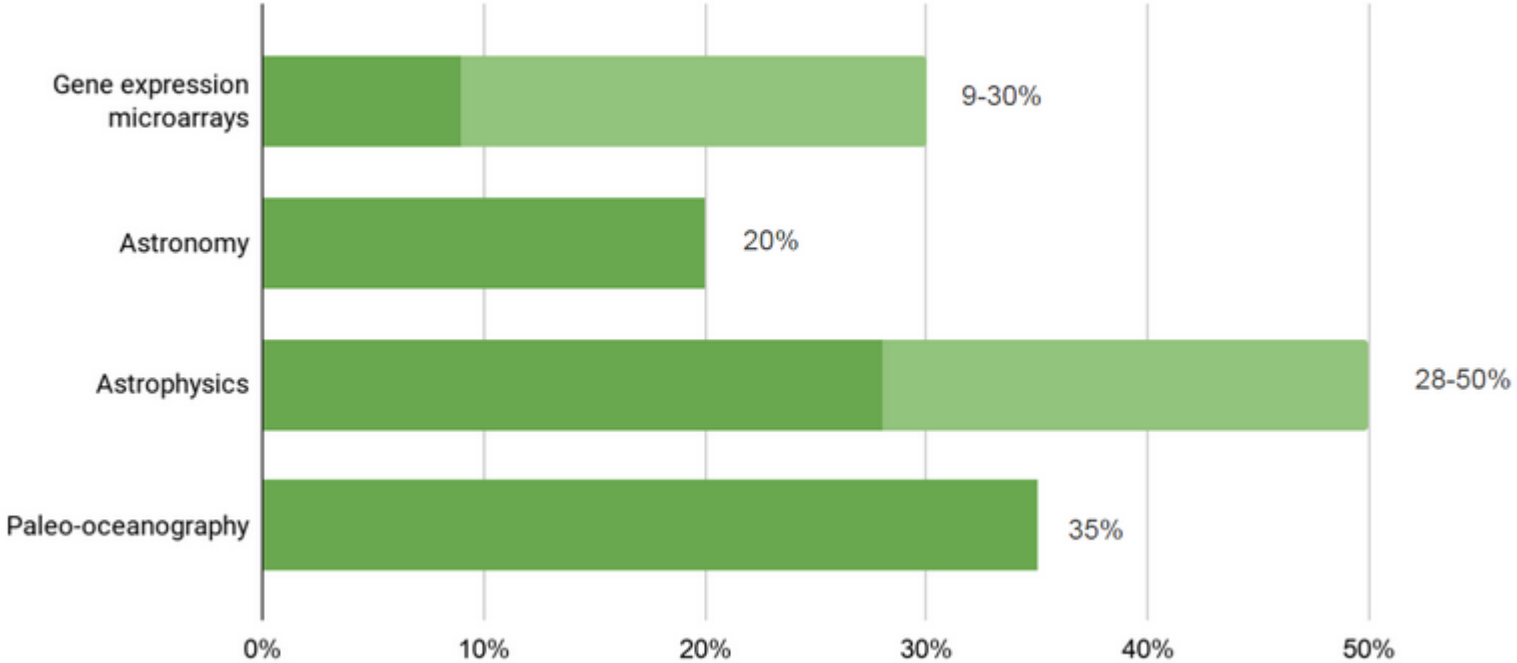
Strategy for archiving



Deposited data – more citations

*We also find an association between articles that include statements that **link to data in a repository** and up to **25.36% ($\pm 1.07\%$) higher citation impact** on average, using a citation prediction model.*

Citation impact of sharing data, by discipline




COMMENTARY

Open Access

Open science saves lives: lessons from the COVID-19 pandemic



Lonni Besaçon^{1,2*} , Nathan Peiffer-Smadja^{3,4}, Corentin Segalas⁵, Haiting Jiang⁶, Paola Masuzzo⁷, Cooper Smout⁷, Eric Billy⁸, Maxime Deforet⁹ and Clémence Leyrat^{5,10}

Abstract

In the last decade Open Science principles have been successfully advocated for and are being slowly adopted in different research communities. In response to the COVID-19 pandemic many publishers and researchers have sped up their adoption of Open Science practices, sometimes embracing them fully and sometimes partially or in a sub-optimal manner. In this article, we express concerns about the violation of some of the Open Science principles and its potential impact on the quality of research output. We provide evidence of the misuses of these principles at different stages of the scientific process. We call for a wider adoption of Open Science practices in the hope that this work will encourage a broader endorsement of Open Science principles and serve as a reminder that science should always be a rigorous process, reliable and transparent, especially in the context of a pandemic where research findings are being translated into practice even more rapidly. We provide all data and scripts at <https://osf.io/renxy/>.

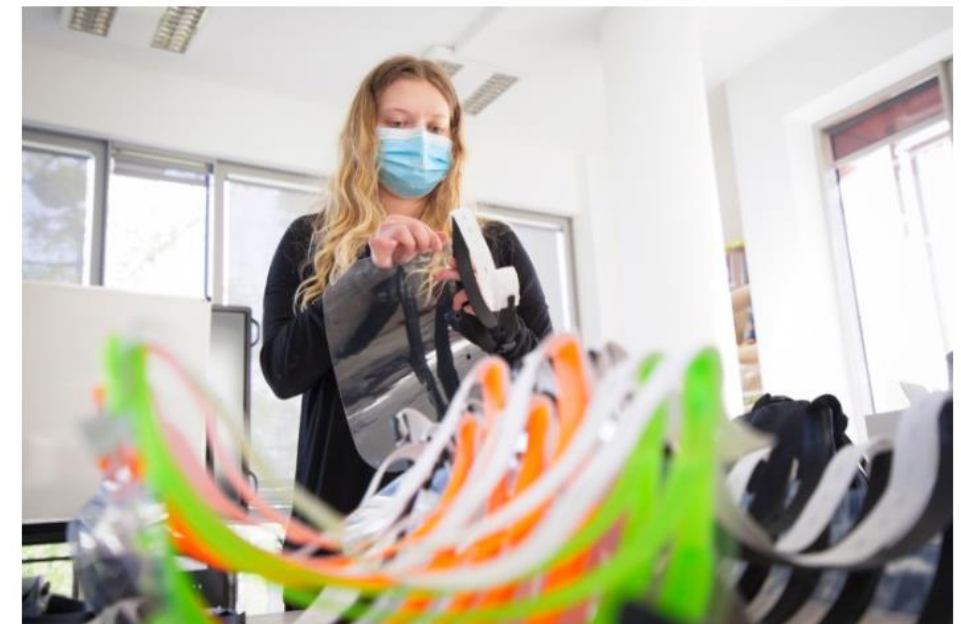
Keywords: Open science, Peer review, Methodology, COVID-19

TECHNOLOGY FEATURE | 24 April 2020

Open science takes on the coronavirus pandemic

Data sharing, open-source designs for medical equipment, and hobbyists are all being harnessed to combat COVID-19.

[Mark Zastrow](#)



A student in Warsaw assembles 3D-printed protective masks. Credit: Jaap Arriens/NurPhoto/Getty

SCIENCE

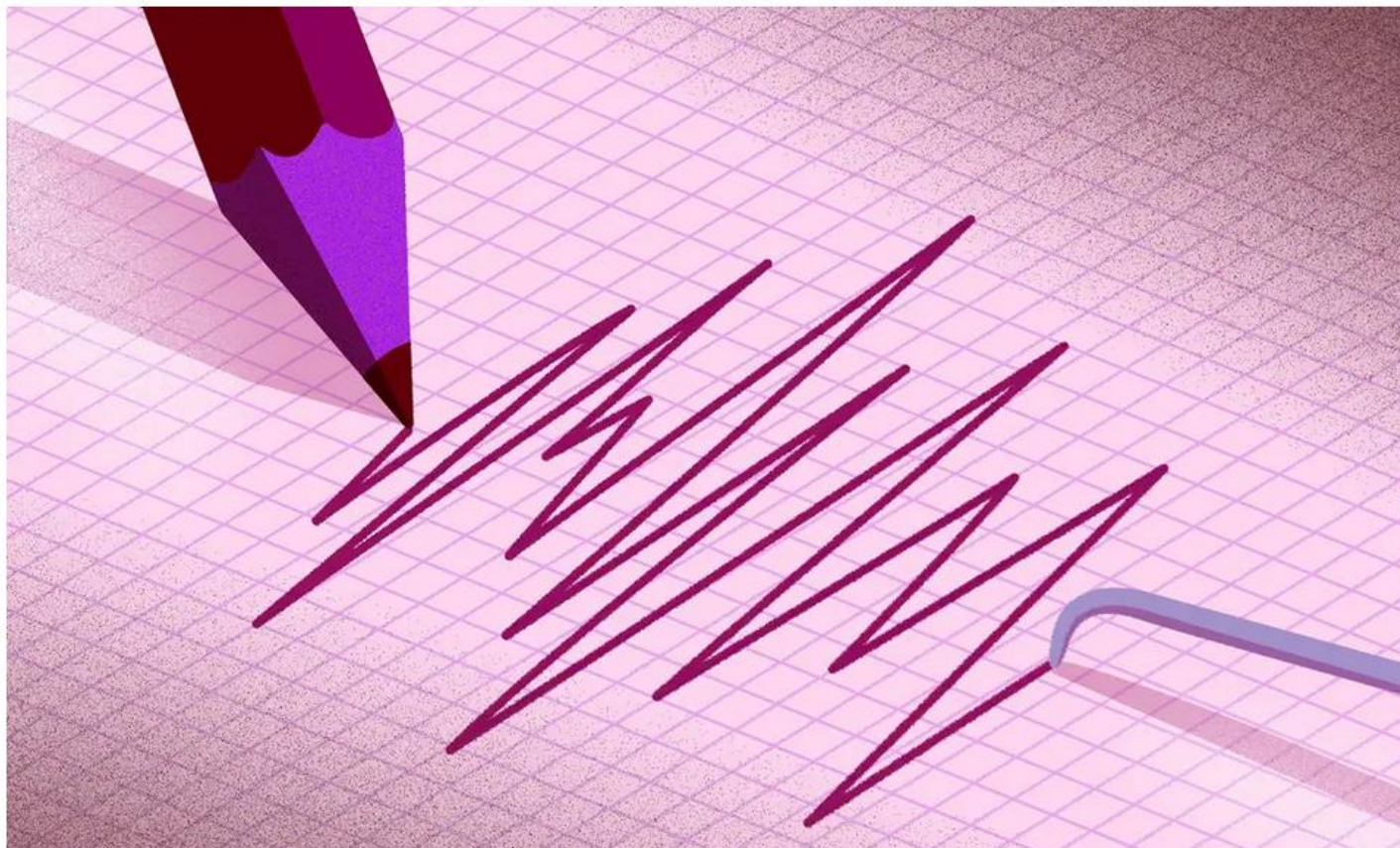
A Famous Honesty Researcher Is Retracting A Study Over Fake Data

Renowned psychologist Dan Ariely literally wrote the book on dishonesty. Now some are questioning whether the scientist himself is being dishonest.



Stephanie M. Lee
BuzzFeed News Reporter

Posted on August 20, 2021, at 2:40 p.m. ET

[Tweet](#)[Share](#)[Copy](#)

"When the researchers published their 2020 update, they posted the data from their 2012 paper for the first time. **Publicly sharing data was once a rarity** in science but is slowly becoming more commonplace amid calls for greater transparency."

Reasons for sharing your data

External Factors

- Funder requirements
- Publisher requirements

Career Benefits

- Increased visibility
- More data reuse
- New collaborations
- Increased citations

Reasons for sharing your data

External Factors

- Funder requirements
- Publisher requirements

Career Benefits

- Increased visibility
- More data reuse
- New collaborations
- Increased citations

Scientific Progress

- More robust research
- Enables verification of results
- Enables new collaborations across disciplines and borders
- Opens up for new uses of data
- Avoids duplication
- Easier to use data in teaching

As open as possible,
as closed as necessary



- Credit must be given to the creator
- Only noncommercial uses of the work are permitted



- Credit must be given to the creator
- Adaptations must be shared under the same terms



- Credit must be given to the creator



- is a public dedication tool, which allows creators to give up their copyright and put their works into the worldwide public domain.

FAIR Data

F
Findable



A
Accessible



I
Interoperable

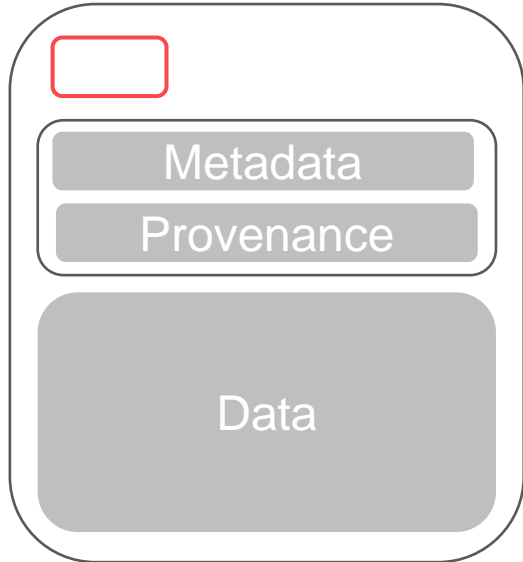


R
Reusable

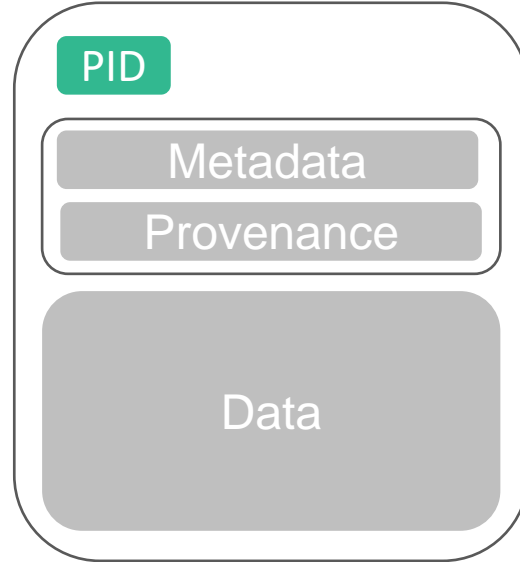


Levels of FAIR

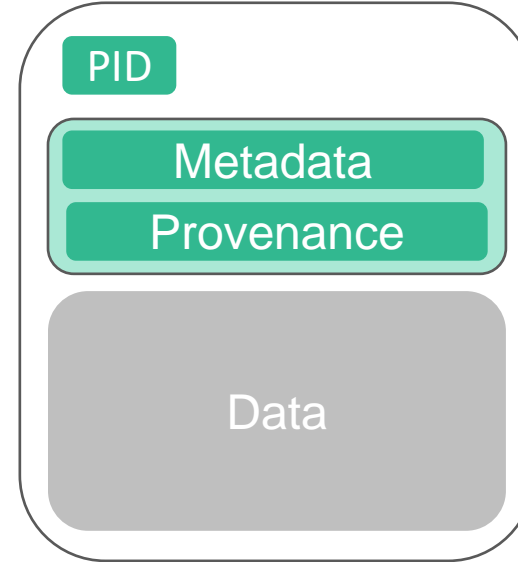
Totally UNFAIR



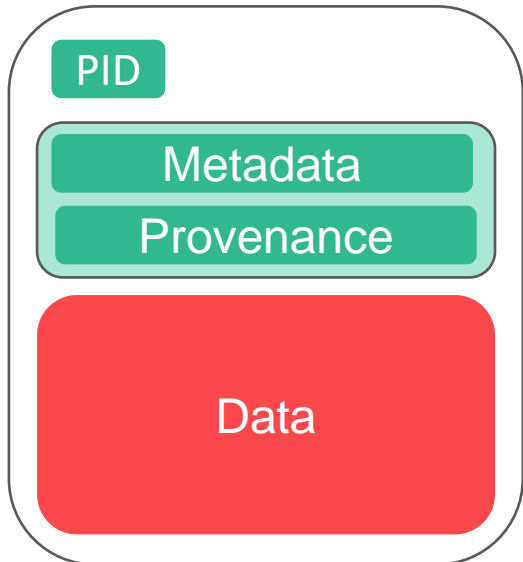
Findable
Usable for humans



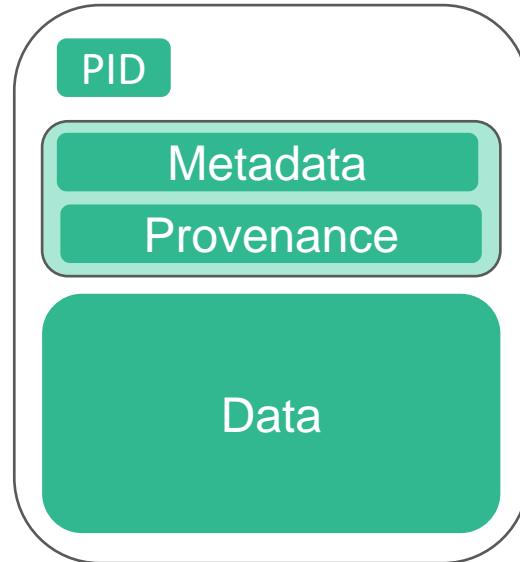
FAIR metadata



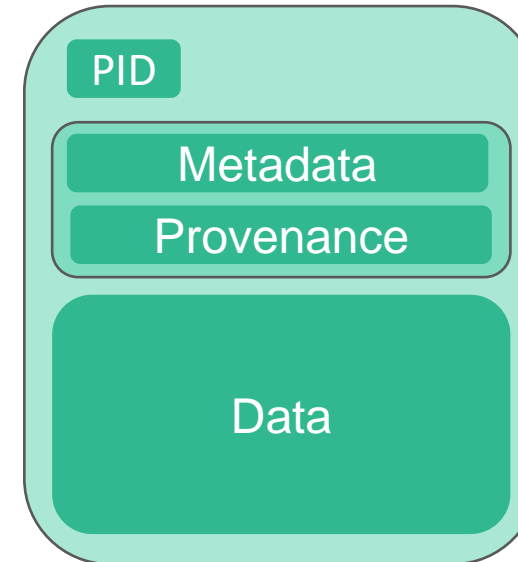
FAIR data
Restricted access



FAIR data
Open access



FAIR data
Open access and functionally linked



Open notebook

Open infrastructures

Open innovation

Open hardware

Crowd funding

Open source

Open Science

Open data

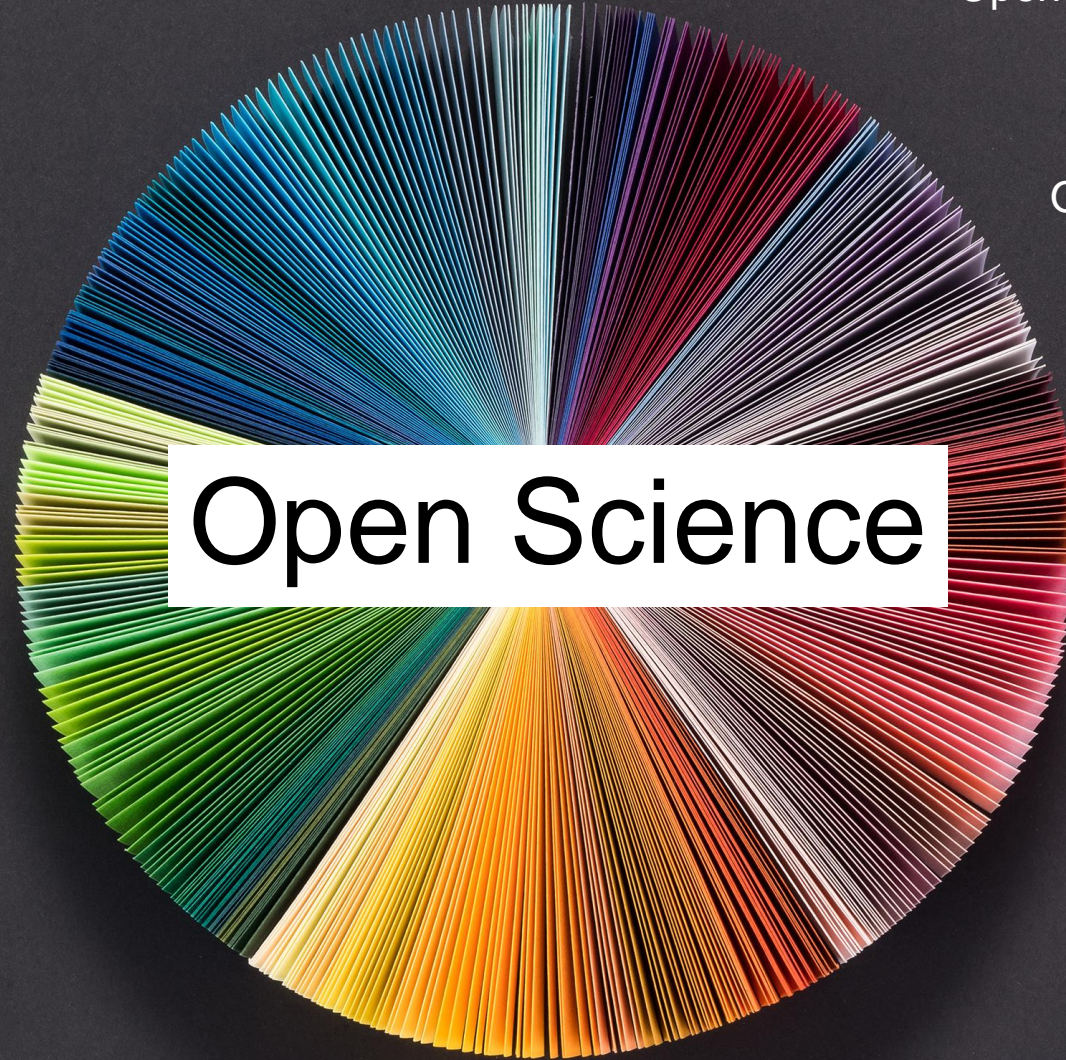
Open labs

Citizen science

Open access

Open evaluation

Open educational resources



Why open science?

improved **quality** of research by **building** on previous works and compiling research data in new ways

transparency in the research process and better opportunity for **verifiability** of scientific results

increased **cooperation** and less duplication of research work

increased **innovation** in the private and public sectors

efficiency improvement and better utilisation of public funds

“My main argument for opening all parts of the process is that is “sharpening” the research process. You cannot be sloppy if you know that it will be exposed”.

Alexander R. Jensenius, 2020



Text documents

Plain text

Markup language

Programming languages

Spreadsheets

Databases

Statistical data

• Preferred format(s)

- PDF/A (.pdf)
- ODT (.odt)

- Unicode text (.txt)

- XML (.xml)
- HTML (.html)
- Related files: .css, .xslt, .js, .es

- MATLAB
- NetCDF
- TextFabric

- ODS (.ods)
- CSV (.csv)

- SQL (.sql)
- SIARD (.siard)
- CSV (.csv)

- SPSS (.dat/.sps)
- STATA (.dat/.DO)
- R

• Non-preferred format(s)

- Microsoft Word (.doc)
- Office Open XML (.docx)
- Rich Text File (.rtf)
- PDF other than PDF/A (.pdf)

- Non-Unicode text (.txt)

- SGML (.sgml)
- Markdown (.md)

- Microsoft Excel (.xls)
- Office Open XML Workbook (.xlsx)
- PDF/A (.pdf)

- Microsoft Access (.mdb, .accdb)
- dBase (.dbf)
- HDF5 (.hdf5, .he5, .h5)

- SPSS Portable (.por)
- SPSS (.sav)
- STATA (.dta)
- SAS (.7dat; .sd2; .tpt)

Materials developed as a part of the *Skills development for research data* project:
<https://www.ub.uio.no/english/about/projects/rdm-skills/>

Contact us at research-data@uio.no





part 2

examples and menti

Links:

FAIR: <https://www.force11.org/fairprinciples>

DataverseNO: <https://dataverse.no>

NSD: <https://www.nsd.no/en/archiving-research-data/>

NIRD: <https://www.sigma2.no/research-data-archive>

Zenodo: <https://zenodo.org>

Figshare: <https://figshare.com>

OSF <https://osf.io>

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

Re3data:

Creative Commons Licenses: <https://creativecommons.org>

CoreTrustSeal:

UiO Research data management <https://www.uio.no/english/for-employees/support/research/research-data-management/>