School on Synchrotron Light Sources and their Applications

Open Science and FAIR data

25th January, 2023

Author: Andy Götz (ESRF)

Role: ESRF Data Policy manager and PaNOSC coordinator

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Outline of Talk

This talk will address the topic of Open Science and FAIR Data for scientists doing research in order to answer the following questions:

- Open Science and FAIR Data

What is this ?
Why do this ?
What to do ?
What to expect ?
What to try ?
What to learn ?

About me

1. Studied **Computer Science** and **Radio Astronomy** in **South Africa + Germany**

2. Joined **ESRF** in **1988**, worked on **accelerator controls, beamline controls, data management**

3. Designed first ESRF control system **TACO (1988-1998)**

4. Leader of team developing the **TANGO** control system **(1999 – now)**

5. Coordinating the **PaNOSC** project on **making FAIR data reality** for Photon and Neutron sources in Europe **(2018 – 2022)**

Kilobytes

to

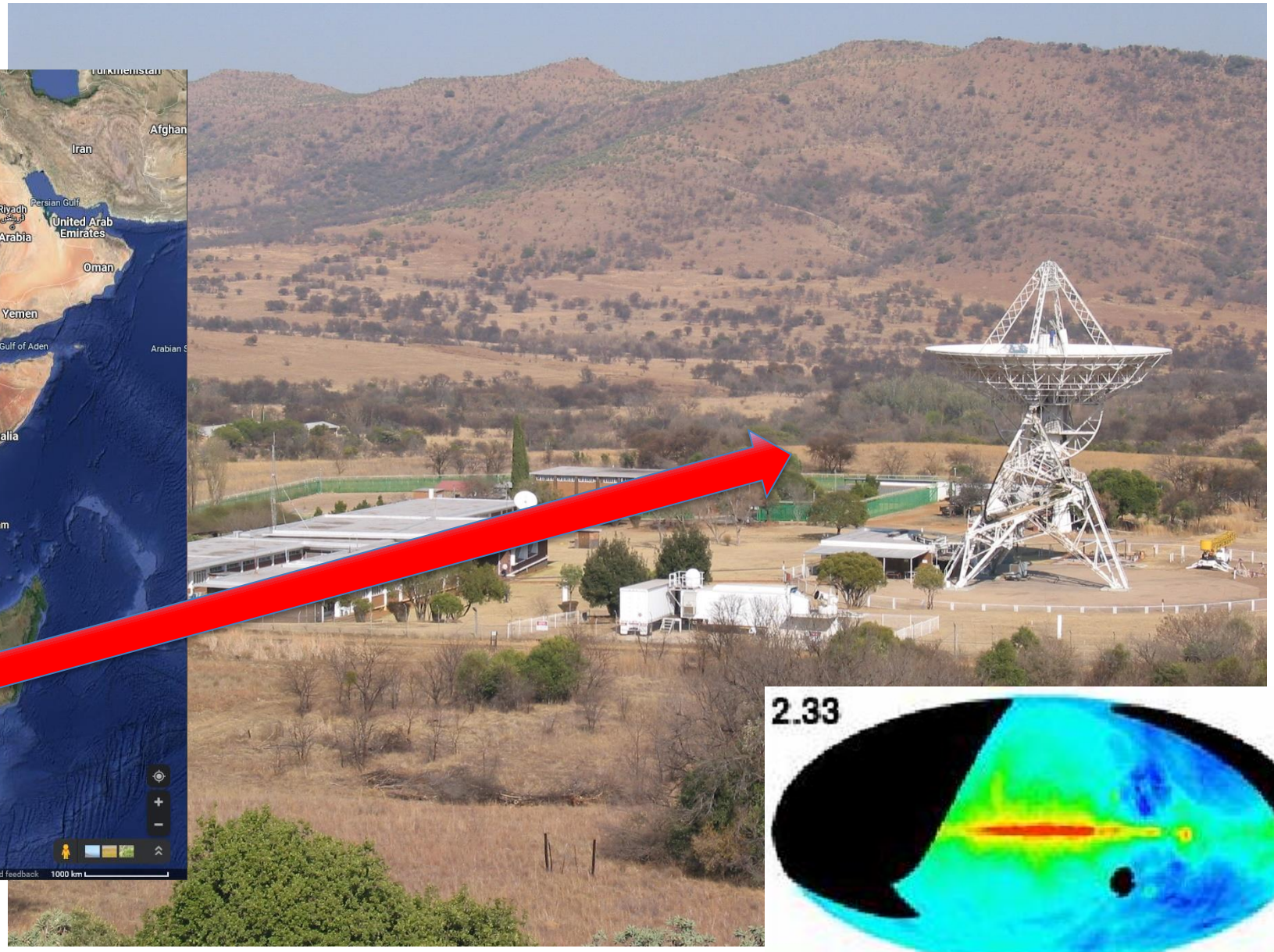
Petabytes

in

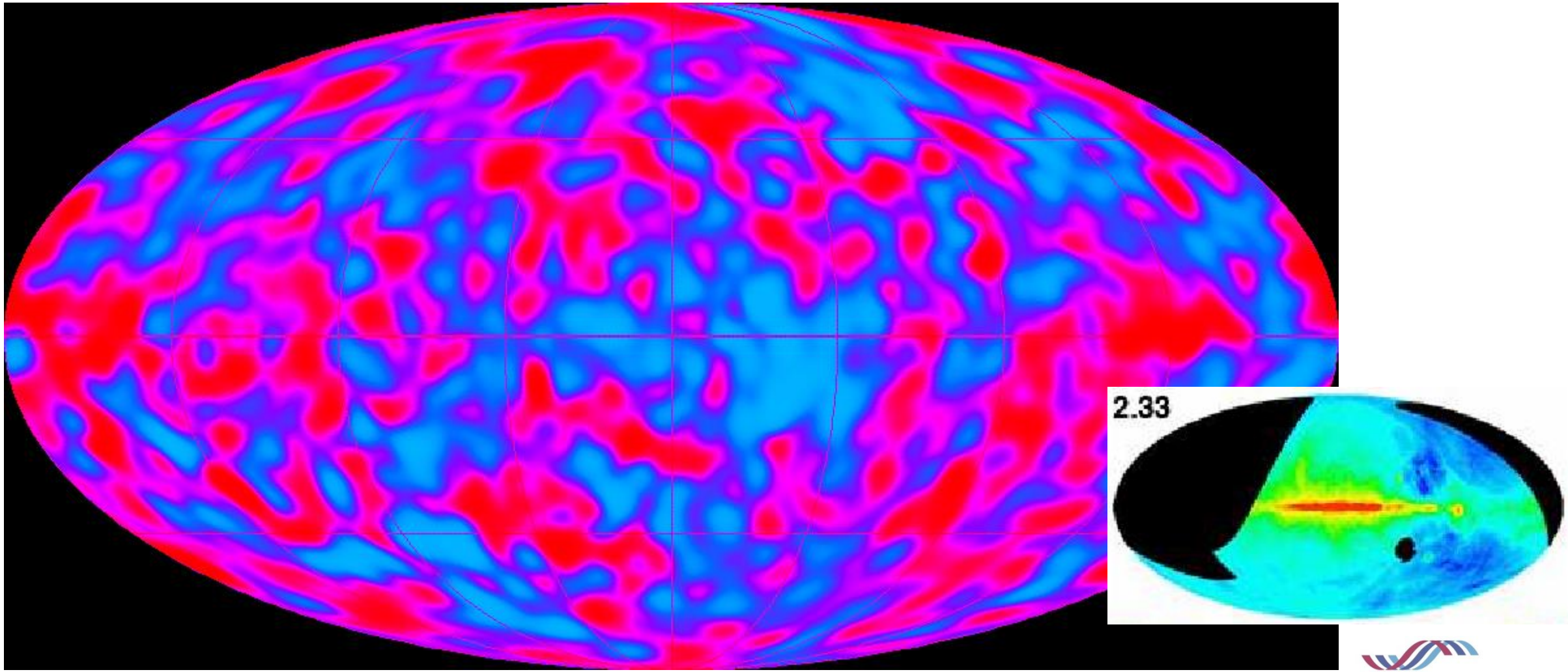
40 years

panosc

Where I started – Hartebeesthoek (South Africa)



CMB anisotropy map formed from data taken by the COBE spacecraft

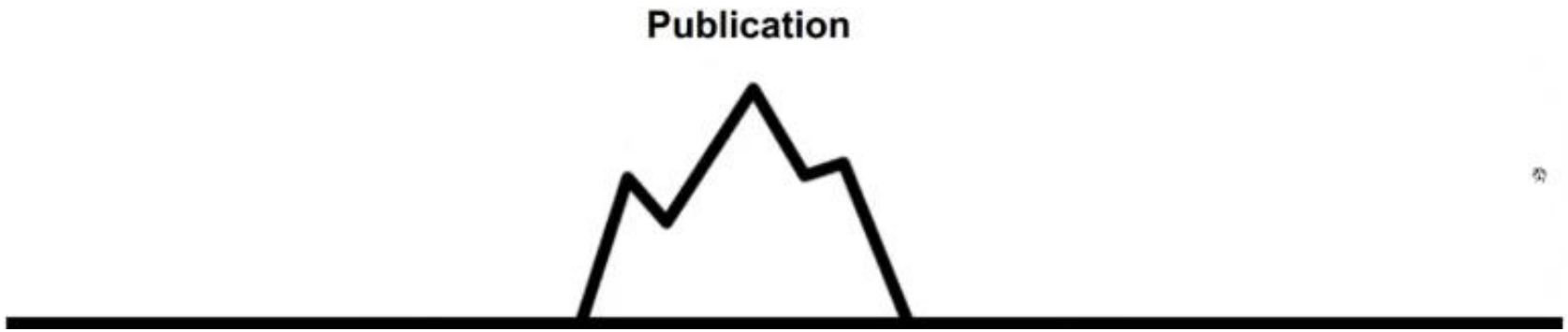


panosc

https://en.wikipedia.org/wiki/Cosmic_Background_Explorer

https://doi.org/10.1111/j.1365-2966.2008.13376.x

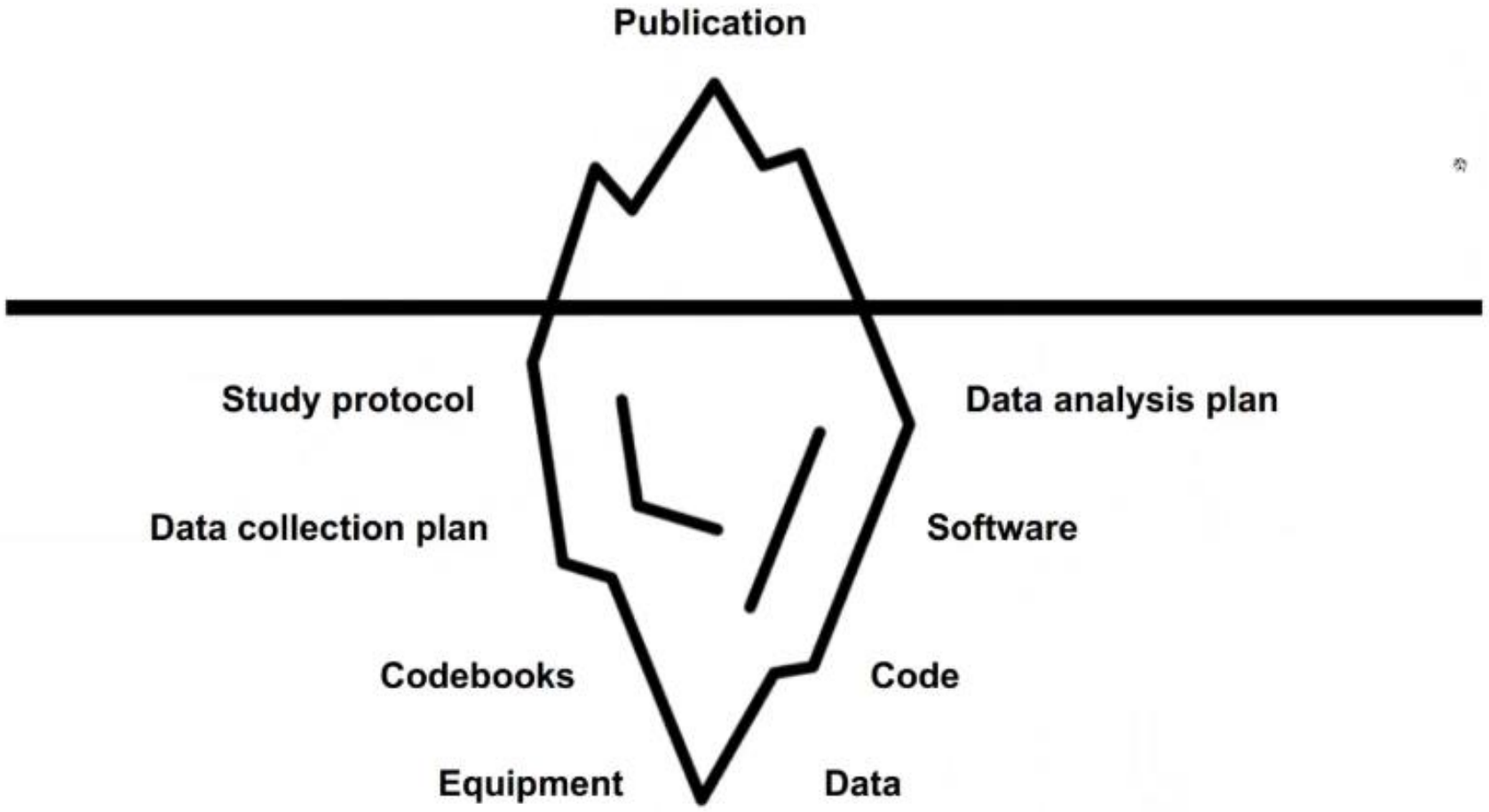
Science produces Publications



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

University of Saskatchewan

Science produces much more than Publications



https://library.usask.ca/studentlearning/workshops/grad-research.php#panel-section-3-ResearchDataManagementWhatYouNeedtoKnow

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

University of Saskatchewan

Reproducibility and Replicability

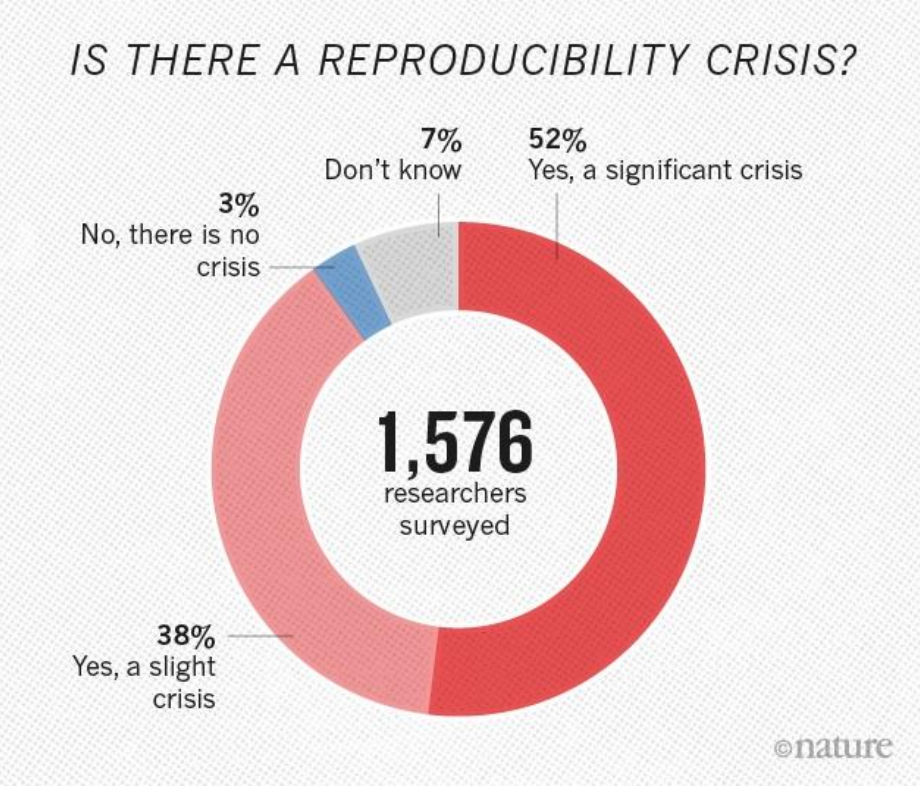
Published: 25 May 2016

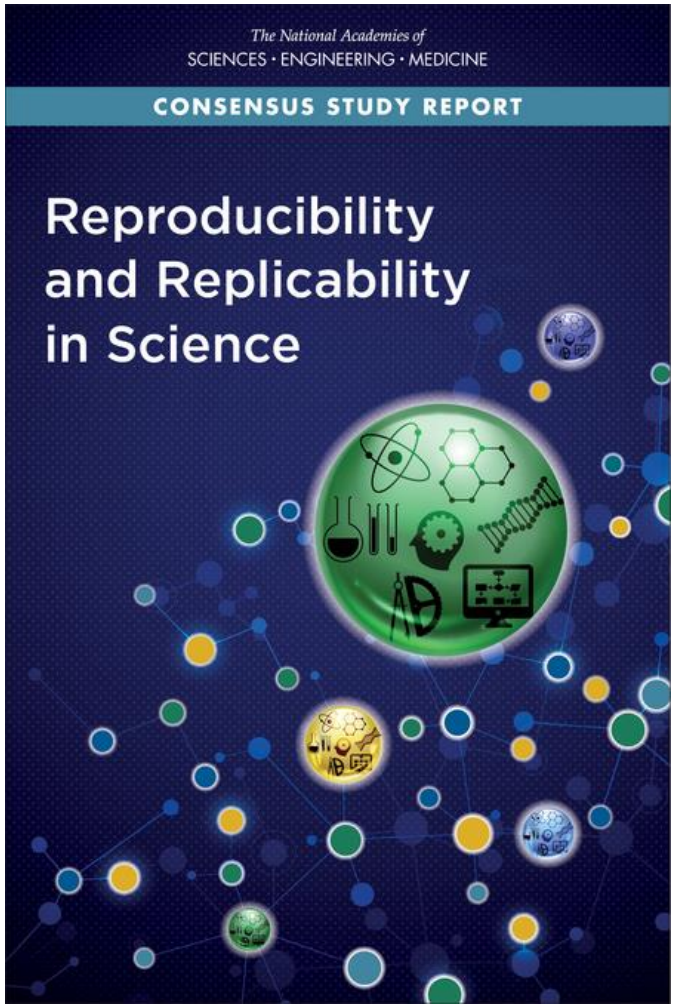
1,500 scientists lift the lid on reproducibility

Monya Baker

Nature **533**, 452–454 (2016) | Cite this article

5320 Accesses | **1225** Citations | **3871** Altmetric | Metrics





Further reading:

- Replication crisis – Wikipedia
- https://phys.org/news/2017-03-science-crisis.html

arch and innovation programme under grant agreement No. 823852

panosc

Open Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

What is Open Science?

- ✓ Open Science as: *"to make the primary outputs of publicly funded research results – publications and the research data – publicly accessible in digital format with no or minimal restriction"*

https://www.oecd.org/sti/inno/open-science.htm

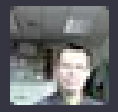
"Work, Finish, Publish" → "**Work, Finish, Publish, Release**"

Michael Farady 1791-1867 Open Science

https://openscience.org/what-exactly-is-open-science/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ChatGPT: What is Open Science?



what is open science

Open science is an umbrella term that encompasses a variety of practices and principles aimed at making scientific research more transparent, accessible, and reproducible. This includes making research data, methods, and findings openly available to the public, as well as promoting collaboration and openness in the scientific process. The goal of open science is to increase the efficiency and effectiveness of scientific research, and to ensure that the knowledge generated by the scientific enterprise is widely accessible and can be used to benefit society. Open science practices can include open data sharing, open access publishing, preregistration of study designs, and the use of open-source software.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

https://chat.openai.com/chat

Pillars of Open Science

1. **Open Access**
   - publications should be freely accessible either as Gold (journal) or Green (preprint) access
2. **Open Data**
   - data should be FAIR and freely accessible under a licence which allows re-use without restriction
3. **Open Source Software**
   - source code should be made available on a publicly accessible repository under an Open Source licence
4. **Open Hardware**
   - hardware designs should be accessible, like software, under an Open Source licence
5. **Open Educational Resources**
   - educational resources (videos, e-training courses etc.) should be made available to all
6. **Citizen Science**
   - citizens who follow the scientific method should be encouraged and facilitated and engage with scientists

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Open Access publications – Green vs. Gold

GREEN

- Articles are free to read after an embargo period
- Bioscientifica automatically make the final published version, also known as the version of record, free
- Authors may deposit a version of their accepted manuscript in an online repository after this time
- There is no cost to authors.

GOLD

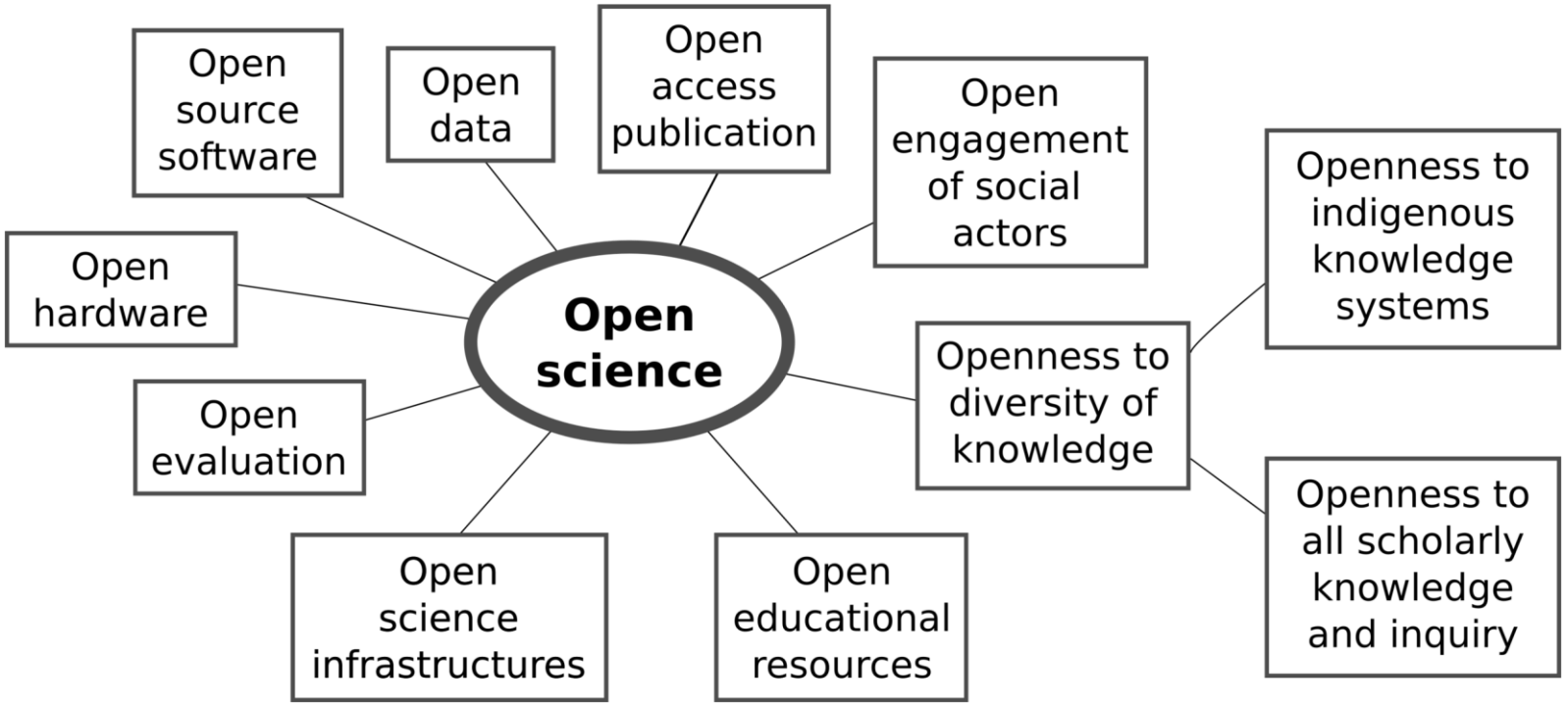
- Authors (or their funders or institutions) pay an Article Publication Charge (APC) upon acceptance
- The final published version is free immediately
- Bioscientifica deposits the article in PubMed Central
- Authors retain copyright and a range of licenses are available
- Journal could be fully open access (eg. *EDM Case Reports*) or hybrid (eg. *European Journal of Endocrinology*).

https://www.bioscientifica.com/authors/preparing-papers/publishing-open-access/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Open Science - origin

"Open Science can be seen as a continuation of, rather than a revolution in, practices begun in the 17th century with the advent of the academic journal, when the societal demand for access to scientific knowledge reached a point at which it became necessary for groups of scientists to share resources with each other" - https://en.wikipedia.org/wiki/Open_science



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Unesco definition of open science

open science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community.

United Nations
Educational, Scientific and
Cultural Organization

Open Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

https://unesdoc.unesco.org/ark:/48223/pf000037

ANNEX VI Recommendation on Open Science

This Recommendation outlines a common definition, shared values, principles and standards for open science *at the international level and proposes a set of actions conducive to a fair and equitable operationalization of open science for all at the individual, institutional, national, regional and international levels.*

https://unesdoc.unesco.org/ark:/48223/pf0000380399

Updated recommendations on the following:

1. Scientific publications
2. Open research data
3. Open educational resources
4. Open source software and source code
5. Open hardware
6. Scientific knowledge
7. Open science infrastructures
8. Open engagement of societal actors
9. Open dialogue with other knowledge systems
10. Public + Private sector

United Nations Educational, Scientific and Cultural Organization

Open Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

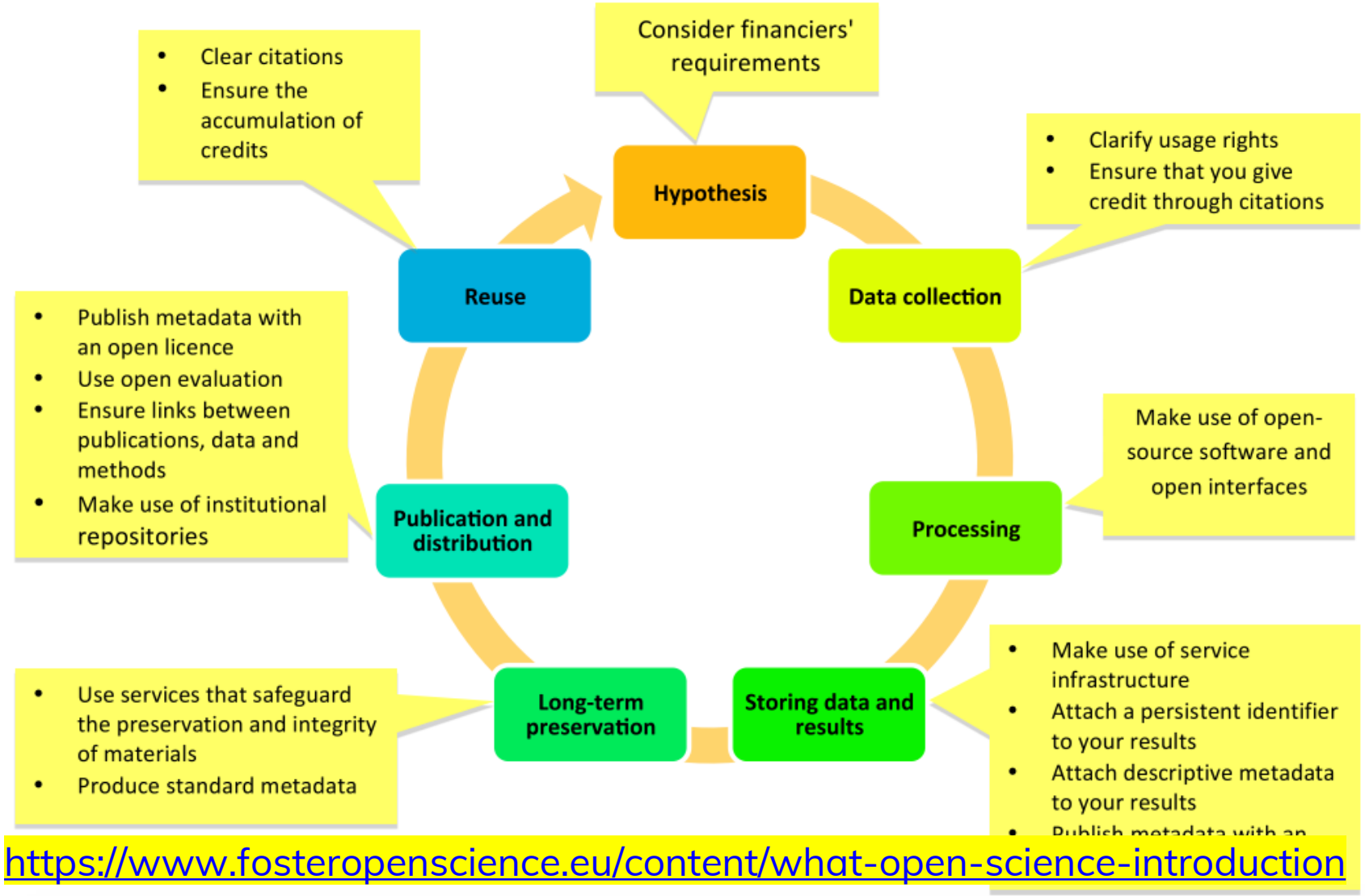
panosc

What is Open science – Open Everything

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852

https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

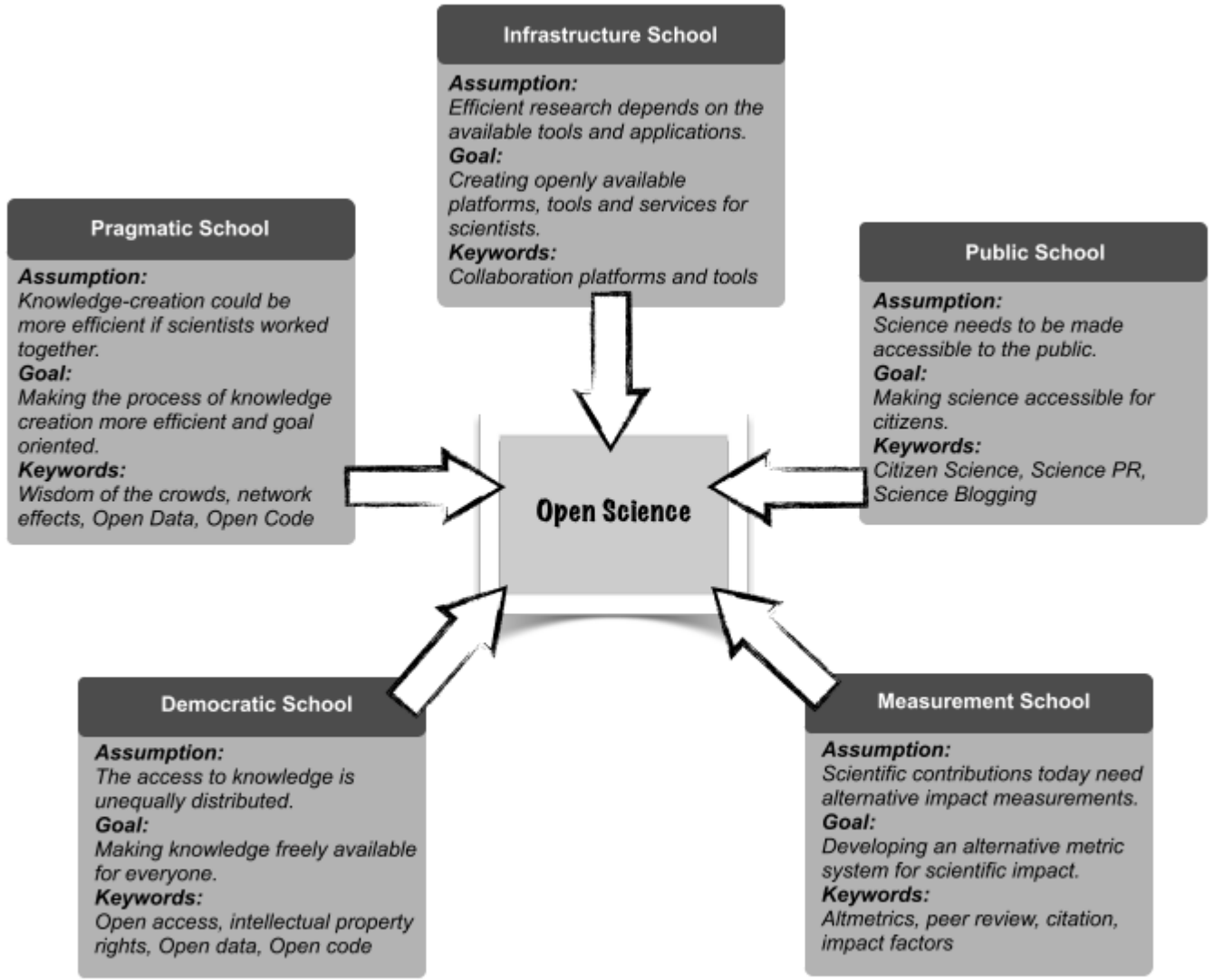
Open Science is about extending the principles of openness to the whole research cycle (FOSTER)



This project has re

https://www.fosteropenscience.eu/content/what-open-science-introduction

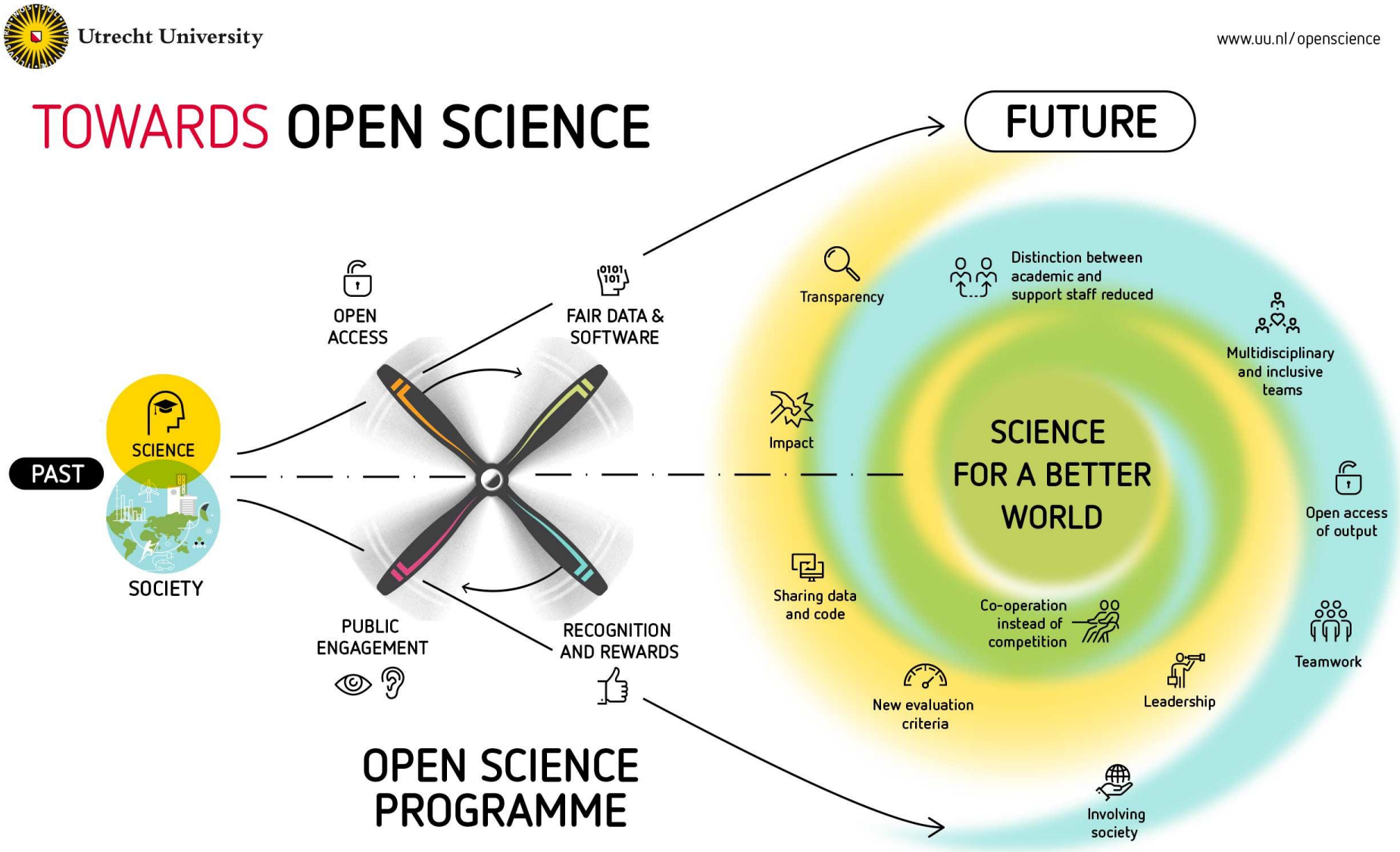
Five schools of Thought for Open Science



This project I

Fecher, B. and Friesike, S. (2013). Open Science: One Term, Five Schools of Thought.

Impact of open science on the future



https://narratives.insidehighered.com/four-pillars-of-open-science/assets/rALfdOrDqh/uugraphic-2-2560x1597.jpeg

European Open Science Cloud

The Vienna Declaration on the European Open Science Cloud

Vienna, 23 November 2018

EU funded 22 EOSC Projects in H2020

We, Minis… …he European…

1. Recall the… …aration" signed in Brussels on 1…

2. Reaffirm t… …ope. Confirm that the vision of t… …nes and Member States, susta…

3. Recognis… …its nature iterative a… …to build trust and con…

4. Hi… …ion of cloud services for S… …he world, reaching out…

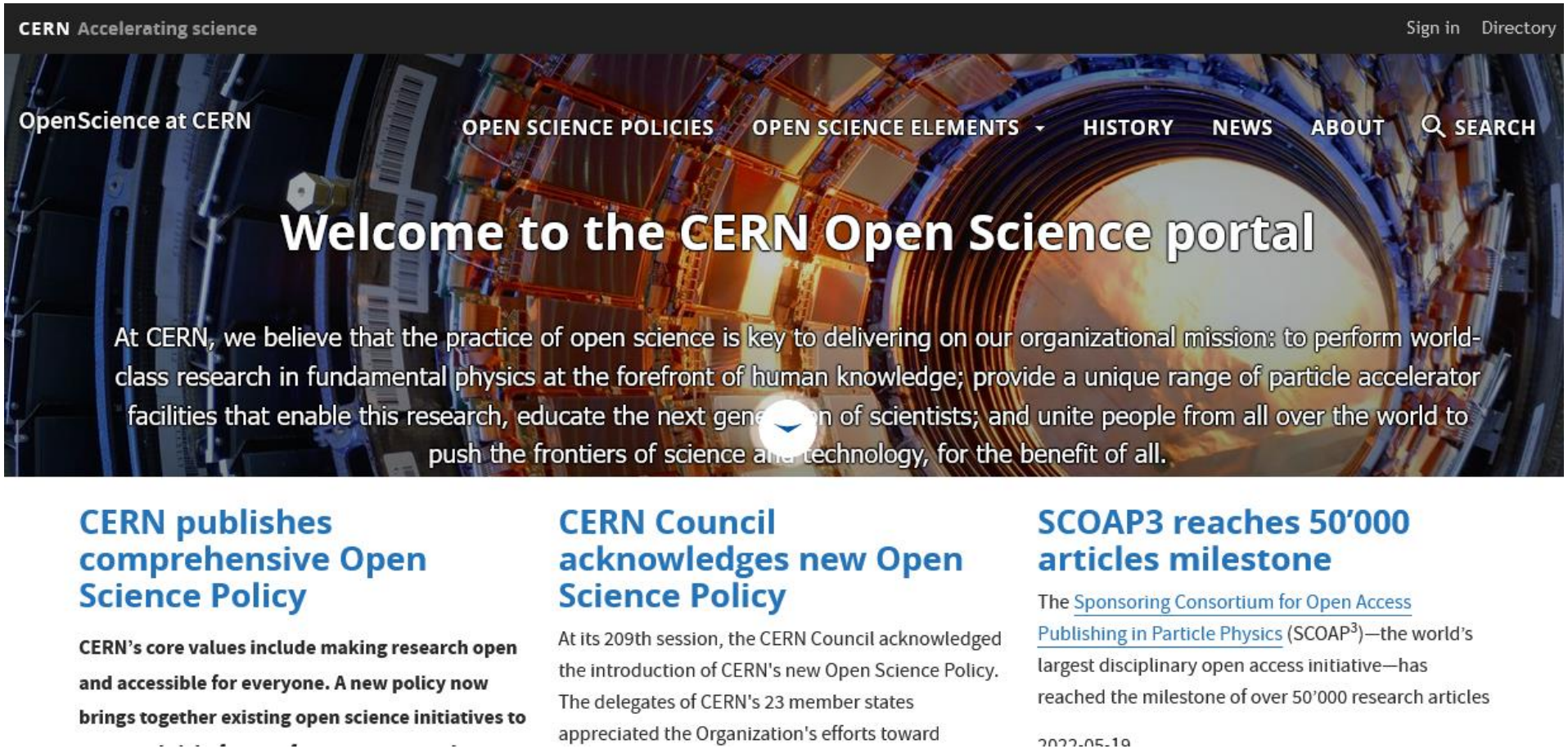
African Open Science Platform

https://aosp.org.za

The Africa Open Science Platform (AOSP) was established in 2017 with an aim to position African scientists at the cutting edge of data intensive science by stimulating interactivity and creating opportunity through the development of efficiencies of scale, building critical mass through shared capacities, and amplifying impact through a commonality of purpose and voice.

CERN publishes Open Science policy

https://openscience.cern/

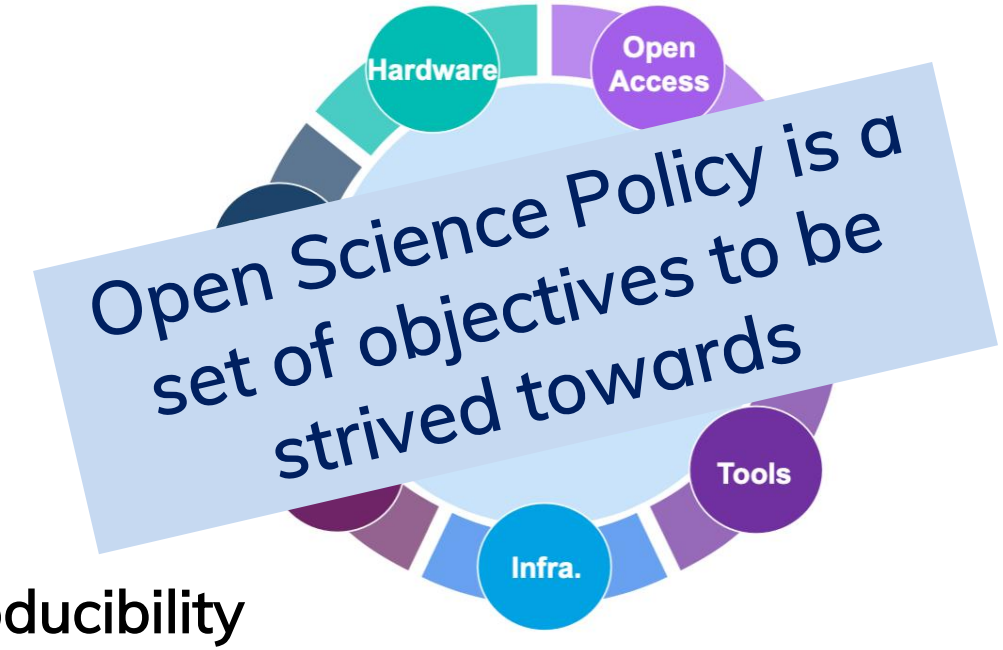


CERN Council adopted an Open Science Policy on 2022-09-29

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Cern Open Science policy

1. Open access to publications
2. Open data
3. Open source software
4. Open hardware
5. Research integrity, reuse and reproducibility
6. Infrastructure provision for open science
7. Research assessment and evaluation
8. Education, training and outreach
9. Citizen Science



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

EMBL Open Science Policy

→ EMBL adopted an Open Science Policy for EMBL staff in December 2021 and is implementing it since January 2022.

→ Two main aspects:

1. Public availability of research outputs
2. Research assessment and fair attribution of credit

1.ORCID

2.DORA

EMBL Open Science Policy. Credit: Holly Joynes/EMBL

https://www.embl.org/documents/wp-content/uploads/2021/12/ip71-open-science-and-open-access-policy.pdf

This project has re… under grant agreement No. 823852

US Federally funded research to be Open Access

BRIEFING ROOM

OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay

AUGUST 25, 2022 • PRESS RELEASES THE WHITE HOUSE

1. *Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;*

2. *Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,*

3. *Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.*

This project has r

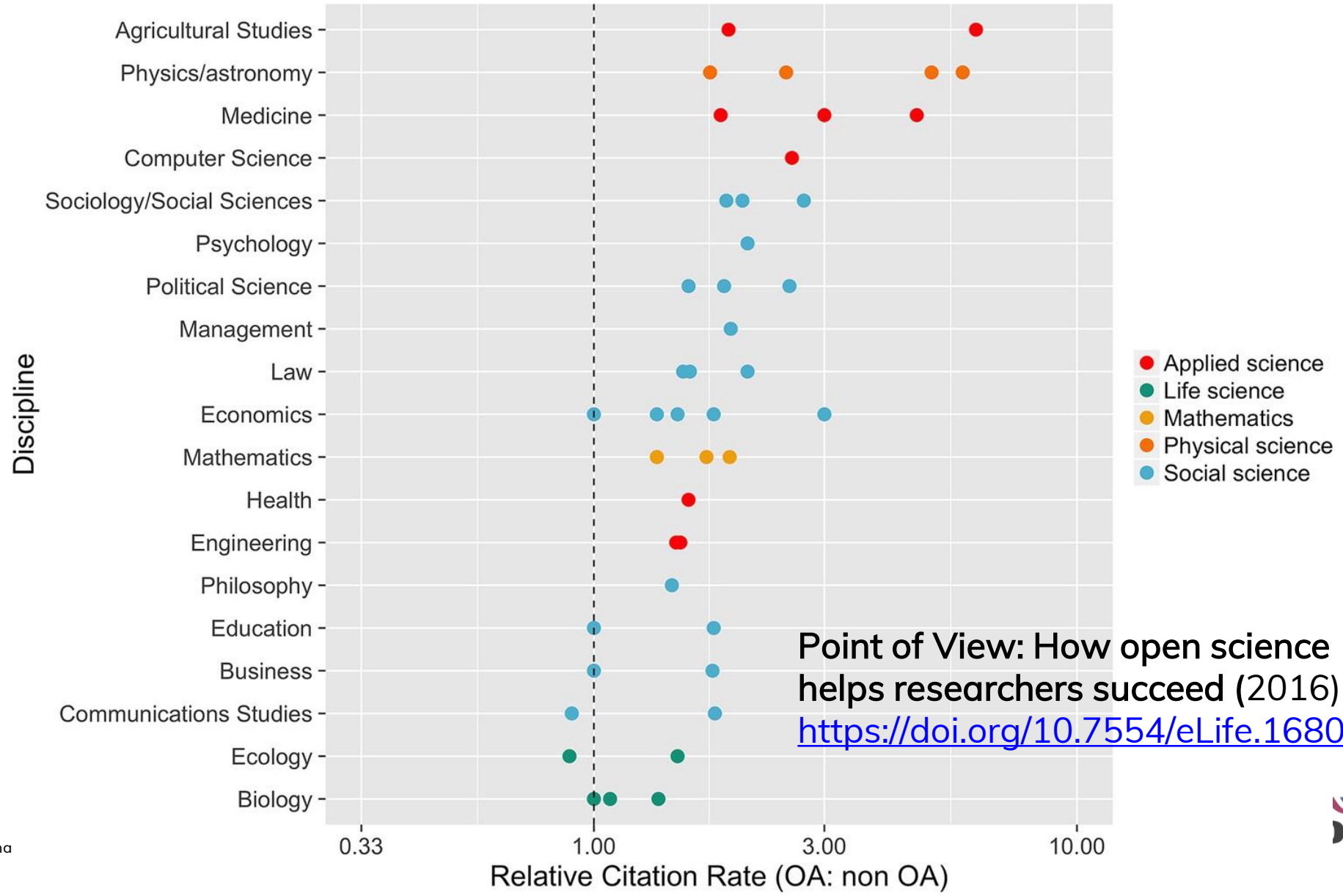
https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf

White House + NASA have declared 2023 Year of Open Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

https://nasa.github.io/Transform-to-Open-Science/year-of-open-science/

Open access leads to more citations



Point of View: How open science helps researchers succeed (2016) https://doi.org/10.7554/eLife.16800

This project ha

Open science is beneficial for scientists

nature > nature methods > articles > article

Article | Open Access | Published: 04 November 2021

Imaging intact human organs with local resolution of cellular structures using hierarchical phase-contrast tomography

C. L. Walsh, P. Tafforeau, W. L. Wagner, D. J. Jafree, A. Bellier, C. Werlein, M. P. Kühnel, E. Boller, S. Walker-Samuel, J. L. Robertus, D. A. Long, J. Jacob, S. Marussi, E. Brown, N. Holroyd, D. D. Jonigk, M. Ackermann & P. D. Lee

Nature Methods **18**, 1532–1541 (2021) | Cite this article

82k Accesses | **25** Citations | **2147** Altmetric | Metrics

This article is in the 99th percentile (**ranked 173rd**) of the 437,805 tracked articles of a similar age in all journals and the 98th percentile (**ranked 1st**) of the 79 tracked articles of a similar age in *Nature Methods*

"If you don't want to share data why become a scientist?"

Claire Walsh (UCL)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agree

https://human-organ-atlas.esrf.eu/

Human Organ Atlas | EXPLORE | SEARCH | RECONSTRUCTIONS | HELP

Welcome to the Human Organ Atlas

Citizen science

The Human Organ Atlas uses **Hierarch** **y** to span a previously poorly explo man anatomy, the micron to

Histology using o other structures with sub-micron accuracy b gan, while clinical CT and MRI scans can image wn to just below a millimetre. HiP-CT bridges these scales i 20 micron voxels, and locally down to microns.

We hope enabled by the ESRF-EBS, will act as a reference to provide new insights int al makeup in health and disease. To stay up to date, follow @HiP-CT

HiP-CT imaging and 3D reconstruction of a complete brain from the body donor LADAF-2020-31. More videos can be viewed on the HiP-CT YouTube channel.

Funding

This project has been made possible by funding from:

- The European Synchrotron Radiation Facility (ESRF) — funding proposal MD-1252
- The Chan Zuckerberg Initiative, a donor-advised fund of the Silicon Valley Community Foundation
- The German Registry of COVID-19 Autopsies (DeRegCOVID), supported by the German Federal Ministry of Health
- The Royal Academy of Engineering, UK
- The UK Medical Research Council

Collaborators

- UCL, London, England: **Peter D Lee, Claire Walsh, Simon Walker-Samuel, Rebecca Shipley, Sebastian Marussi, Joseph Jacob, David Long, Daniyal Jafree, Ryo Torii, Charlotte Hagen**
- ESRF, Grenoble, France: **Paul Tafforeau, Elodie Boller**
- Medizinische Hochschule Hannover, Germany: **Danny D Jonigk, Christopher Werlein, Mark Kuehnel**
- Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Germany:**M Ackermann**
- University Hospital of Heidelberg, Germany: **Willi Wagner**
- Grenoble Alpes University, Department of Anatomy, French National Center for Scientific Research: **A Bellier**

Hierarchical imaging of complete human organs

This project

Open science vs. science

Most of these assumptions are not new, as the tradition of openness itself is at the roots of science, but the current developments of information and communication technologies have transformed the scientific practices to a level that requires a different approach to research (FOSTER)

https://www.fosteropenscience.eu/content/what-open-science-introduction

Q: "What is the difference between Open Science and 'science'?"

A: *Open Science refers to doing traditional science with more transparency involved at various stages, for example by openly sharing code and data. Many researchers do this already, but don't call it Open Science.*

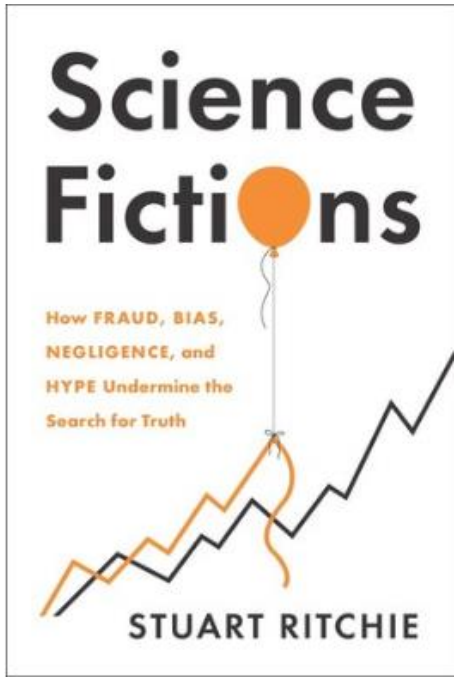
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

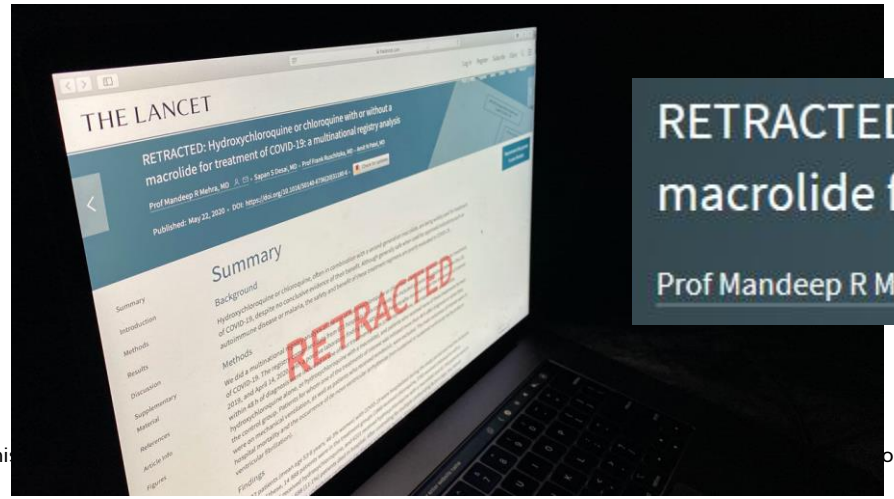
European Conduct of Scientific Integrity

Integrity, scientific method, open science

- Recommend to follow the EU Code of Integrity
  - https://allea.org/code-of-conduct/

- To AVOID having your papers RETRACTED
  - https://retractionwatch.com/





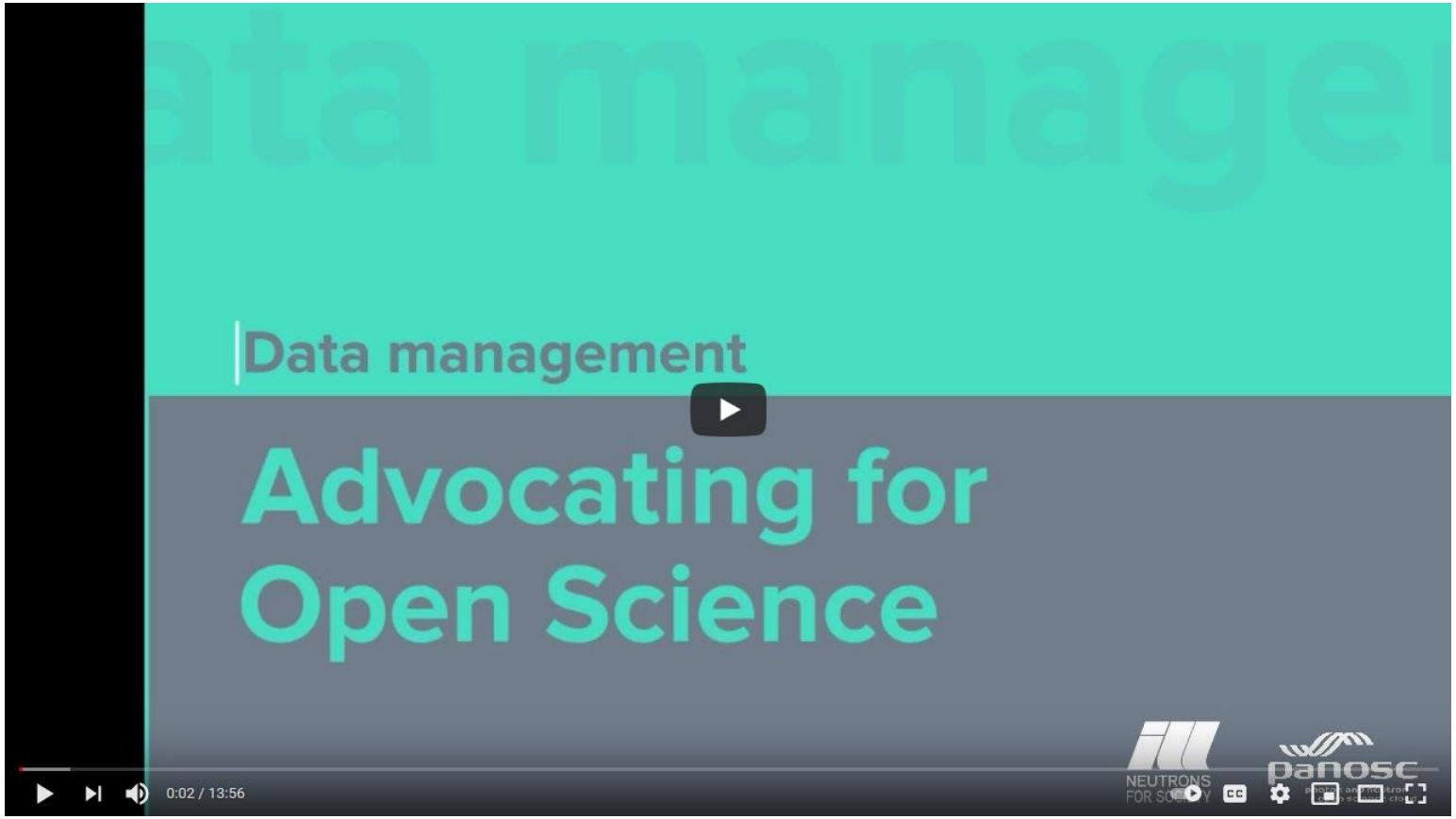
RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Prof Mandeep R Mehra, MD • Sapan S Desai, MD • Prof Frank Ruschitzka, MD • Amit N Patel, MD

This ovation programme under grant agreement No. 823852

Open Science Ambassador

Watch this interview of Petr Čermák, a strong advocate of open on the advantages of Open Science for neutrons and science in general



https://youtu.be/QKAc1y6HZNk

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

https://coara.eu

Coalition for Advancing Research Assessment

'Publish or perish' and metrics have led us into a blind alley. Let's start recognizing the full breadth of value created by researchers.

Marc Schiltz

President of Science Europe

I believe in a research culture that recognises a diversity of contributions to science and society; that celebrates high quality and impactful research; and that values sharing, collaboration, integrity and engagement with society, transmitting knowledge from generation to generation.

Mariya Gabriel

Commissioner for Innovation, Research, Culture, Education and Youth

This project has received fundin

panosc

Further reading – Open Science

Many resources are available on Open Science, here are some used for this talk

- Phys.org
  - Five questions about open science answered
  - Data sharing can offer help in science's reproducibility crisis
- UNESCO
  - Recommendation on Open Science
- EU
  - Progress on Open Science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

FAIR Data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

The publication that started the FAIR movement

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

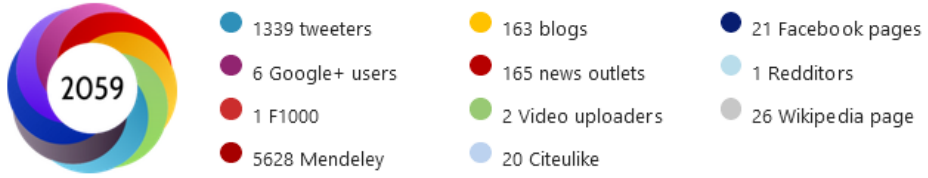
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, ... Barend Mons

+ Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

523k Accesses | **5193** Citations | **2059** Altmetric | Metrics

Online attention



This article is in the 99th percentile (ranked 41st) of the 299,830 tracked articles of a similar age in all journals and the 99th percentile (ranked 1st) of the 23 tracked articles of a similar age in *Scientific Data*

https://data.europa.eu/doi/10.2777/1524

https://www.go-fair.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data availability – the wrong + right way

Data availability

Data available on reasonable request to the authors.

Open Research

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6993871 , reference number 6993871.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

FAIR Principles

https://www.go-fair.org/fair-principles/

Findable

- F1: (Meta) data are assigned globally unique and persistent identifiers
- F2: Data are described with rich metadata
- F3: Metadata clearly and explicitly include the identifier of the data they describe
- F4: (Meta)data are registered or indexed in a searchable resource

Accessible

- A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
- A1.1: The protocol is open, free and universally implementable
- A1.2: The protocol allows for an authentication and authorisation where necessary
- A2: Metadata should be accessible even when the data is no longer available

Interoperable

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

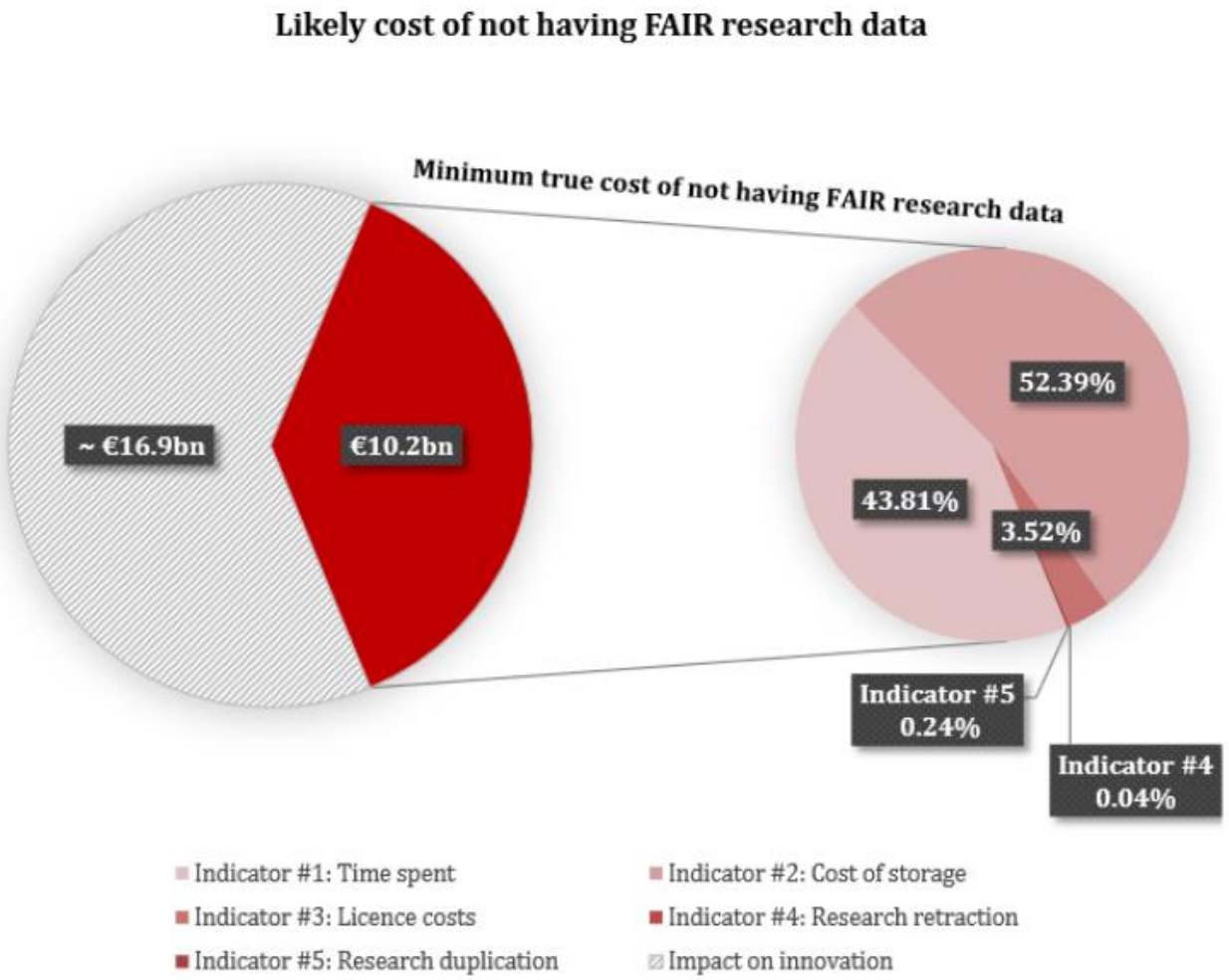
Reusable

- 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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

The cost of not having FAIR data = estimated €10.2bn / year

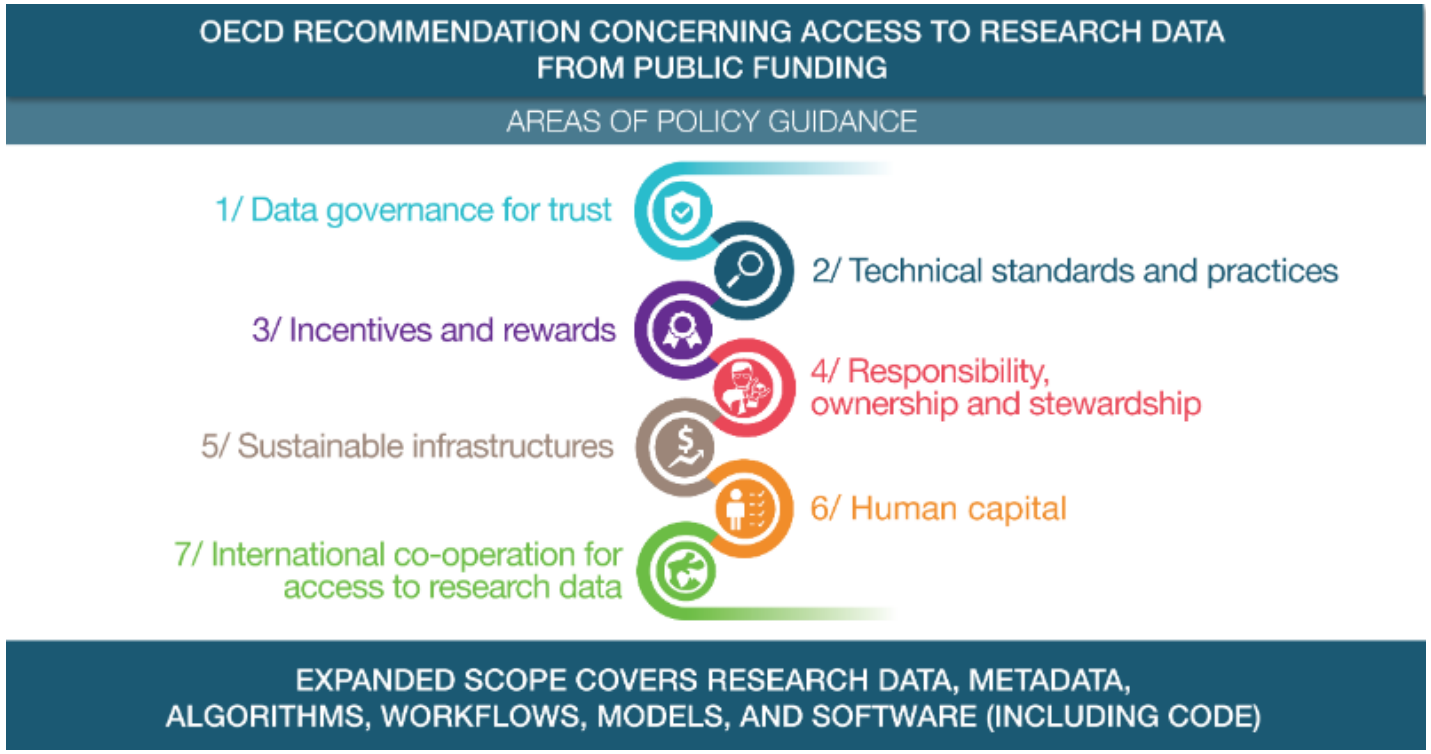


"Cost-benefit analysis for FAIR research data " (https://op.europa.eu/s/pevt)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Open data for publicly funded research

- The **OECD** recommendation in **2006** had a big impact on data policies
- The recommendation was updated in **2021** (https://www.oecd.org/sti/recommendation-access-to-research-data-from-public-funding.htm)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data policies

1. **Check the research-data requirements of your funding agency and field of research.**

A Data policy defines the rules of access and usage to the data produced. Research Institutes like the EIROforum ones all have data policies in place now.

- You are required to accept the data policy when requesting access
- Data is not considered as property but has a usage licence
- Data are under **embargo** (varying from 1 yr, 3 yr, 5 yr) for use by the original creators for a limited amount of time **before being made open**.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ERIOforum member Data Policies

- **CERN** – open data policy for LHC (since 2020)
- **EMBL** – open access policy (since 2015)
- **ESA** – open data policy for most data (since 2010)
- **ESO** – open data policy (updated in 2016)
- **ESRF** – open data policy (since 2015)
- **EUROfusion** – proposal for open data policy (in progress since 2018)
- **EuXFEL** – open data policy (since 2017)
- **ILL** – open data policy (since 2012)
- Others
  - **CERIC-ERIC** – open data policy (since 2021)
  - ....

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

CERN announces new open data policy in support of open science

11 December 2020.

A new open data policy for scientific experiments at the Large Hadron Collider (LHC) will make scientific research more reproducible, accessible, and collaborative

- *The four main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have unanimously endorsed a new **open data policy** for scientific experiments at the **Large Hadron Collider** (LHC), which was presented to the CERN Council today. The policy commits to **publicly releasing** so-called **level 3 scientific data**, the type required to make scientific studies, collected by the LHC experiments. Data will start to be released approximately **five years after collection**, and the aim is for the **full dataset** to be publicly available by the close of the experiment concerned. The policy addresses the growing movement of **open science**, which **aims to make scientific research more reproducible, accessible, and collaborative**.*

https://home.cern/news/press-release/knowledge-sharing/cern-announces-new-open-data-policy-support-open-science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ESRF Data Policy

https://www.esrf.fr/datapolicy

ESRF EUROPEAN SYNCHROTRON RADIATION FACILITY

30 November 2015

The ESRF Data Policy

The ESRF aims to implement a Data Policy starting as soon as possible in 2016. The main elements of this policy comprise:

- **Data ownership**
- **Data curation**
- **Data archiving**
- **Open access to data**

This policy follows largely the recommendations of the PaN-data Europe Strategic Working Group laying out a common framework for scientific data management at photon and neutron facilities (Deliverable D2.1, PaN-data Europe, co-funded by the European Commission under the 7th Framework Programme)

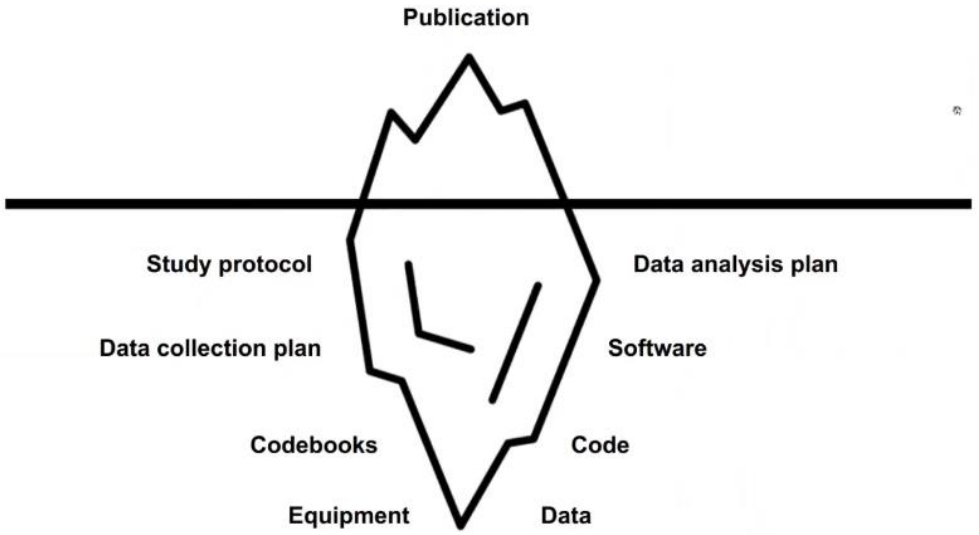
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

Data and research outputs

3. **List the various types of data and research outputs that you expect to produce.**

- Output from your research is everything you produced to come up with your findings including :
  - Raw data
  - Metadata
  - Processed data
  - Analysis workflows
  - Logbooks
  - Software
  - Etc.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Metadata and Why it is important

8. **Provide metadata that allows others to understand, cite and reuse your data files.**

Documentation or information about a data set.

https://data.research.cornell.edu/content/writing-metadata

- **Metadata is all additional data you need to understand your data**

- Examples range from file name, time, to experiment condition, energy, sample name, sample parameters, …

- Use the standard vocabularies defined for your domain e.g. Nexus, FITS, …

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Metadata vocabularies

Many standard vocabularies exist for processed data. There are fewer vocabularies for raw data but they do exist. Check the existing standards for your domain.

- Don’t invent a new vocabulary until you are sure none exists

- Databases of standard vocabularies:
  - https://fairsharing.org/ - FAIRsharing as a community approach to standards, repositories and policies
  - https://www.dcc.ac.uk/guidance/standards/metadata/list - list of Metadata standards

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

Metadata – Take away messages

Metadata have a tendency to get treated as 2nd class data.

Whatever you do TAKE YOUR METADATA SERIOUSLY !

The quality of your data depends on it!

- RECORD them DIGITALLY
- STORE them with your DATA
- FOLLOW the STANDARD(s)
- ENSURE others can UNDERSTAND your (meta)data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Example vocabulary – Nexus for photon and neutron sources

NeXus

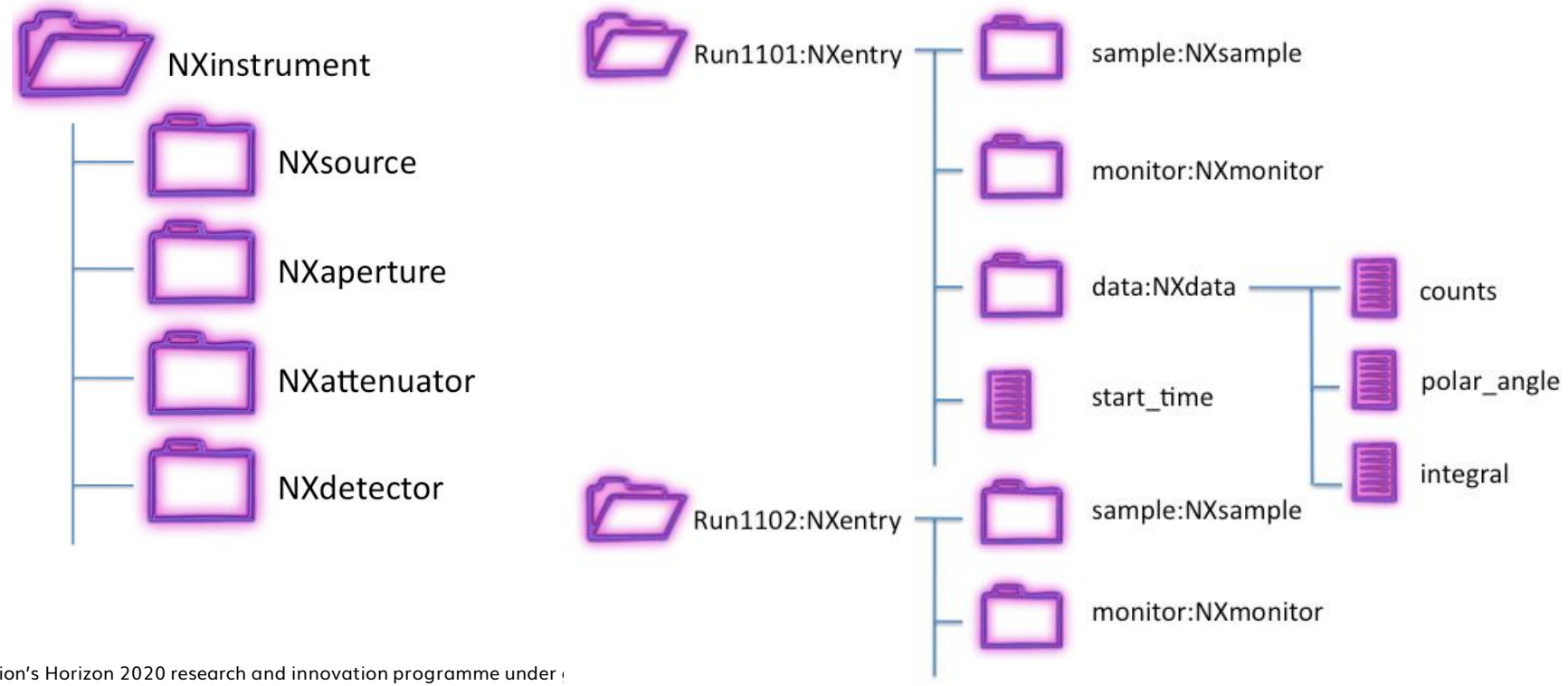
NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.

Home

GitHub Organisation

https://www.nexusformat.org/

Nexus provides a standard vocabulary for:



European Union's Horizon 2020 research and innovation programme under

Example vocabulary – Nexus for photon and neutron sources

Example of structure of data file from ESRF:

Name	Description	Type	Shape	Link
lima.h5		NXroot		
entry_0000	"Lima 2D de...	NXentry		
end_time	"2020-09-08...	string	scalar	
instrument		NXinstrument		
mpx_cdte_22_eh1		NXdetector		
acquisition		NXcollection		
data	3D data	uint16	100 × 516 × 516	
detector_information		NXcollection		
header		NXcollection		
image_operation		NXcollection		
plot		NXdata		
data	3D data	uint16	100 × 516 × 516 Soft	
measurement		NXcollection		
data	3D data	uint16	100 × 516 × 516 Soft	
start_time	"2020-09-08...	string	scalar	
title	"Lima 2D de...	string	scalar	

NeXus

NeXus is developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.

Home

GitHub Organisation

© 2021 NIAC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

Data formats

5. Define appropriate data file formats (see https://fairsharing.org/ for formats).

7. Check what data format and structure the chosen archive might request.

Data formats refer to how the bytes in a file are interpreted. Not the data vocabularies. Data formats must be readable over the long term (for archiving). Data formats must be efficient

- Example data formats:
  - CSV (Comma Separated Values)
  - TIFF for images
  - HDF5 as container
- USE the STANDARD(s) for your community

Further reading: ETD Guidance Brief File Formats

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

E-logbooks

Provide metadata that allows others to understand your experiment.

Logbooks are an essential part of the scientific method. All scientists should keep a logbook. E-logbooks replace paper logbooks.

- E-logbook advantages
  - Shared editing online
  - Powerful search facilities
  - Access rules during embargo period
  - Allows others to understand what you did during the experiment
- **E-logbook is metadata** and will be part of the open data

Further reading: https://guides.library.oregonstate.edu/research-data-services/data-management-lab-notebooks

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ESRF e-logbook example – ID21 / EV-280

Data Portal | My Data | Open Data | Closed Data | Shipping | Manager | Log out Andy GÖTZ

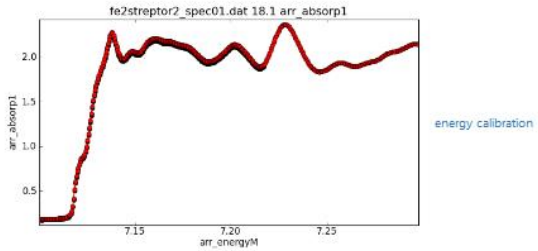
ID21 / EV-280 Beneficial symbiosis in tomato plants: its role on Fe translocation and speciation

Dataset List 90 | Logbook | Shipping

+ New | Take a photo | View | PDF | Help | 0 new log(s) arrived. | Everywhere | Everywhere | 72 found

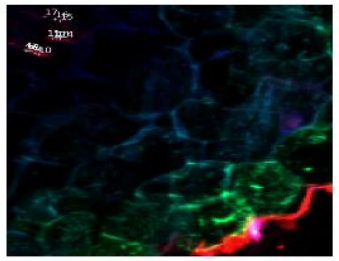
November 5th 2018

10:18:40 `OPTICS> zapenergy 7.1 7.3 400 0.1 1 2 0 2000 (zap: #2, spec: #19)`



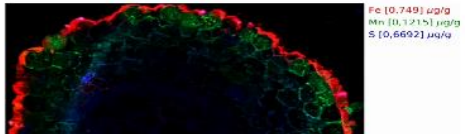
energy calibration

09:58:09 `OPTICS> Fexanes_ev280`

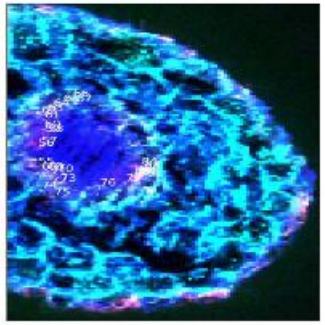


Fe2 strepto r2 Main Root: XANES Points

00:18:27 `OPTICS> zapxiaimage samy 7.904 7.324 580 samz 25.476 26.025 549 100 0 (zap: #1, spec: #2)`

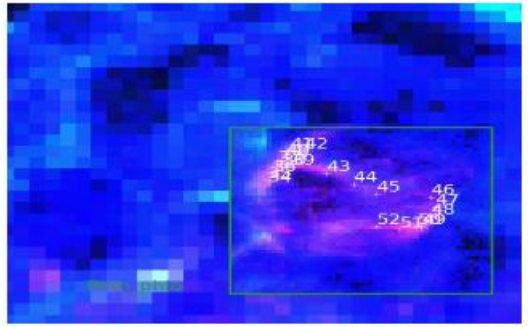


23:50:29 `OPTICS> Fexanes_ev280`



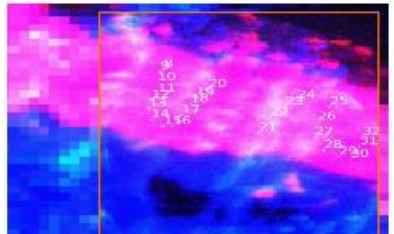
Fe3 prio r3 Sec Root: XANES Points

22:48:31 `OPTICS> Fexanes_ev280`

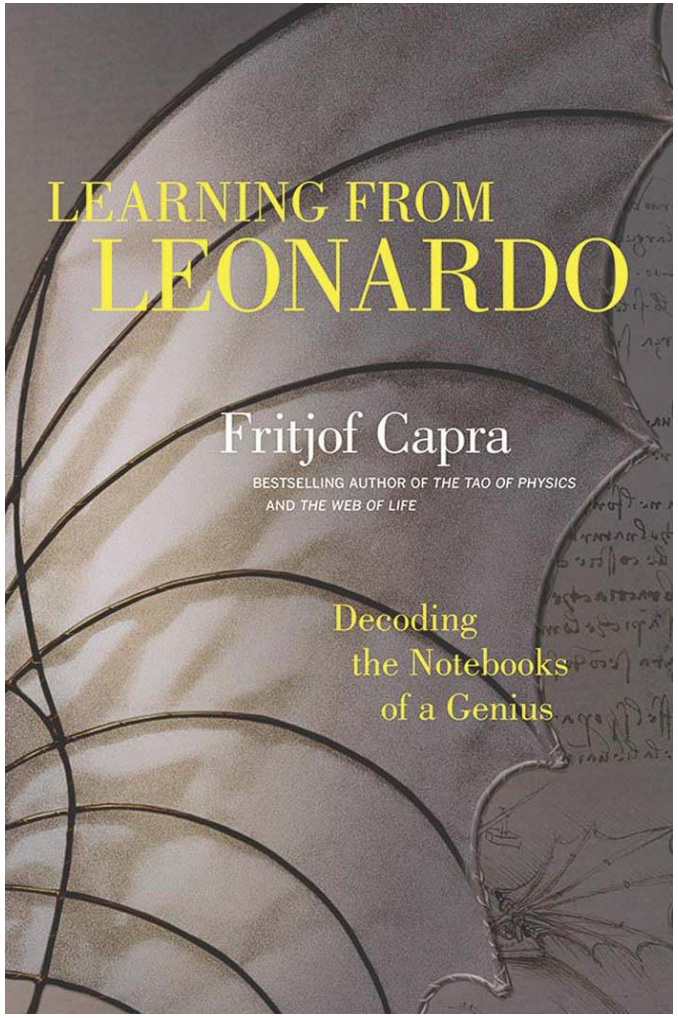


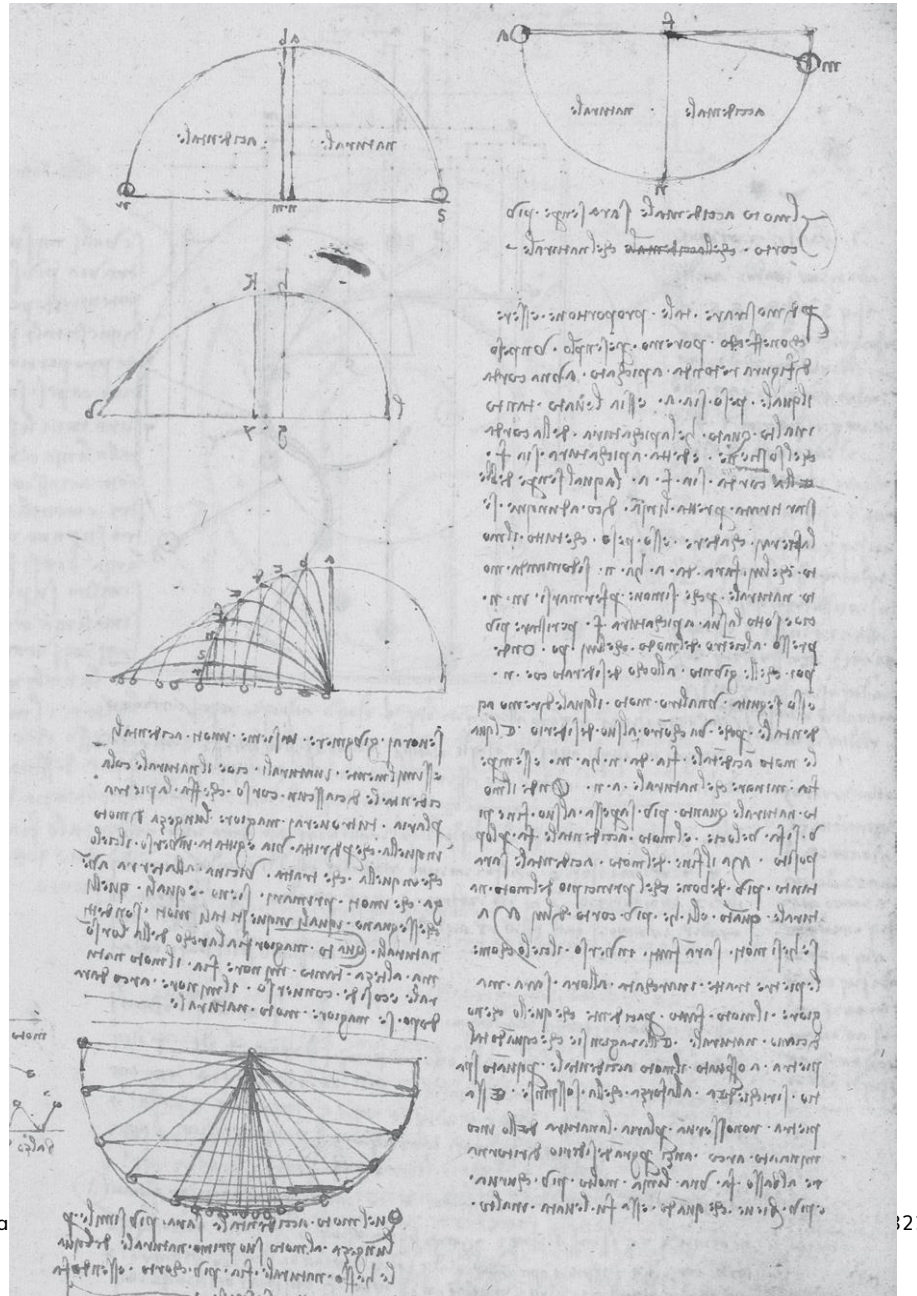
Fe3 prio r3 Main Root prim phlo: XANES Points

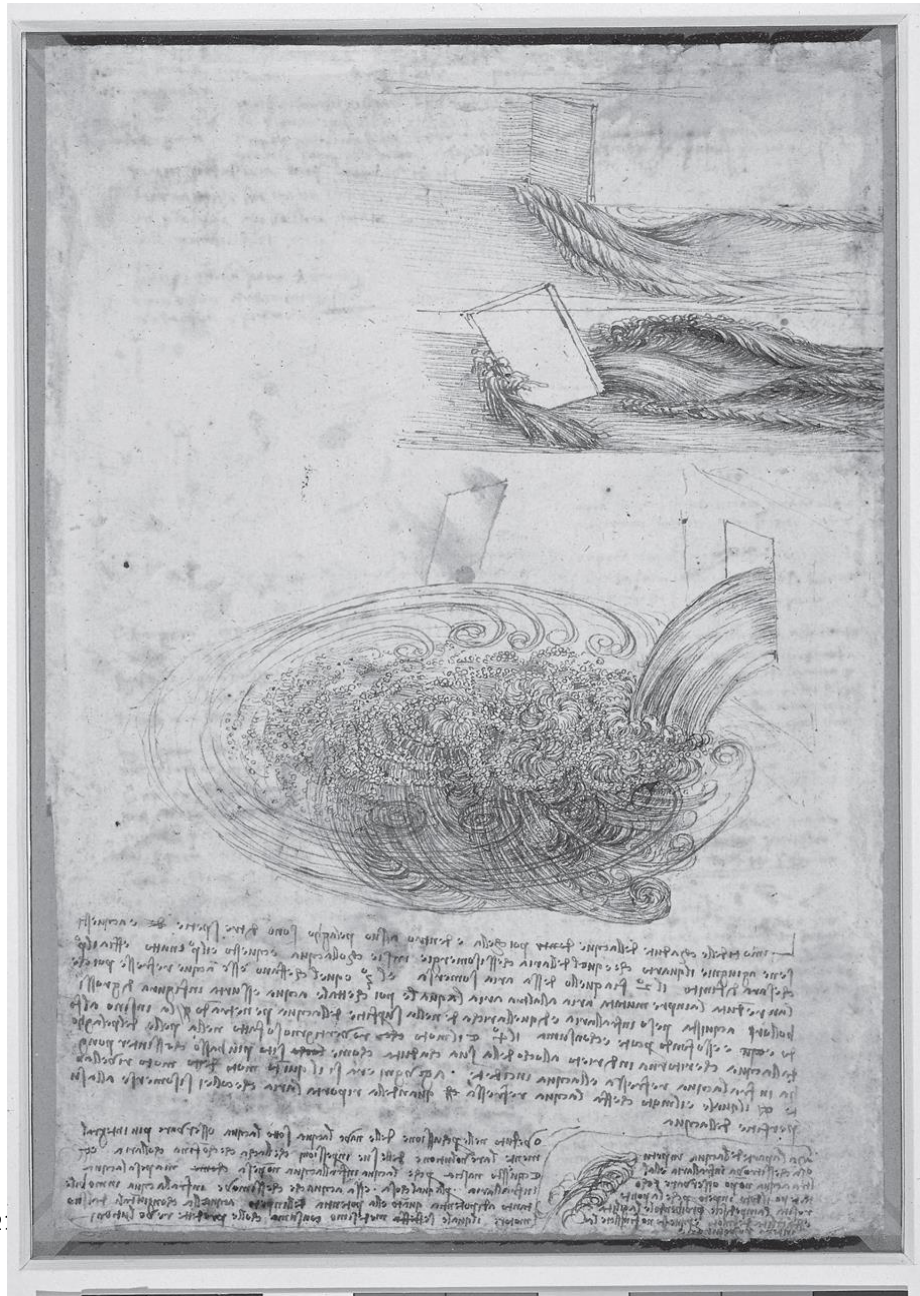
22:25:02 `OPTICS> Fexanes_ev280`



Notebooks can inspire Logbooks e.g. **Leonardo da vinci's** notebooks







notebooks can be very useful for posterity…

This project has received funding from the Europea

Open Source Software

Software is an essential part of a scientists toolset. Many scientists have learned to program so they can analyse their data. The resulting software is part of the outcomes of the research.

- Wherever possible **use Open Source software**

- When **writing software** :
  - Follow **best practices** for software
  - Publish it under an **Open Source license**
  - Store it in an **open (Git) repository** with **version control**
- **Cite your software** in your publications

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

E-Life author guide

https://reviewer.elifesciences.org/author-guide/full

- Source Code:
    - *Relevant software or source code should be deposited in an open software archive. Where appropriate, authors can upload source code files to the submission system (for example, MATLAB, R, Python, C, C++, Java).* ***Any code provided should be properly documented, in line with these instructions*** *(courtesy of PLOS). Please also refer to our Software sharing policy.*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Software tools

Many specific and generic tools exist. One common tool which is being adopted widely is JupyterLab and the Python language.

- **Python** has become the de facto programming language in science
- **Jupyter** notebooks enable reproducible publications https://jupyter.org
- **Binder** service can preserve and run the software for an analysis - https://mybinder.org/

Jun 2021	Jun 2020	Change	Programming Language	Ratings	Change
1	1		C	12.54%	-4.65%
2	3	▲	Python	11.84%	+3.48%
3	2	▼	Java	11.54%	-4.56%
4	4		C++	7.36%	+1.41%
5	5		C#	4.33%	-0.40%
6	6		Visual Basic	4.01%	-0.68%
7	7		JavaScript	2.33%	+0.06%

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

Data Management Plans (DMP)

2. **Go online for help in developing a data-management plan. A useful guide outlining UK funder expectations can be found at go.nature.com/2tnohla.**

12. **Revisit your plan frequently and update it if necessary.**

- DMP document the data management steps in a more formal manner
- Funders are requiring DMPs to ensure RDM is planned
- Facilities will require DMPs more and more to be sure Users can deal with the research data
- DMPs are living documents which need to be updated throughout the project
- Examples of DMPs can be found on DMPonline

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Typical questions to be answered by the DMP

- What data will be created during research.
- Which policies might apply to the data, such as legal, institutional and funding requirements.
- Which data standards will be used, including metadata standards.
- How data will be documented.
- Ownership, copyright and intellectual property rights in data.
- Data security aspects.
- Data storage and backup measures and required equipment or infrastructure.
- Plans for sharing data, who will have access and whether there are any embargoes or restrictions.
- Data management roles and responsibilities.
- Costing or resources needed over and above usual research and dissemination activities to enable data sharing (certainly for the shorter term following the end of any funded research project).

"Managing and Sharing Research Data: A Guide to Good Practice" by Louise Corti et al

https://study.sagepub.com/corti2e

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data repositories

6. **Look for data repositories used by your research community or your host institution (see www.re3data.org for examples).**

A data repository stores data for citing, accessing and archiving data over the long term. Repositories can be provided by facilities or community based. Choose the right repository with the service you expect

- Facilities offer repositories for raw and (sometimes) processed data e.g. https://data.esrf.fr
- Choose repository which is certified e.g. http://go.nature.com/2eLHBFP)
- Use an institute or community archive which is sustainable

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

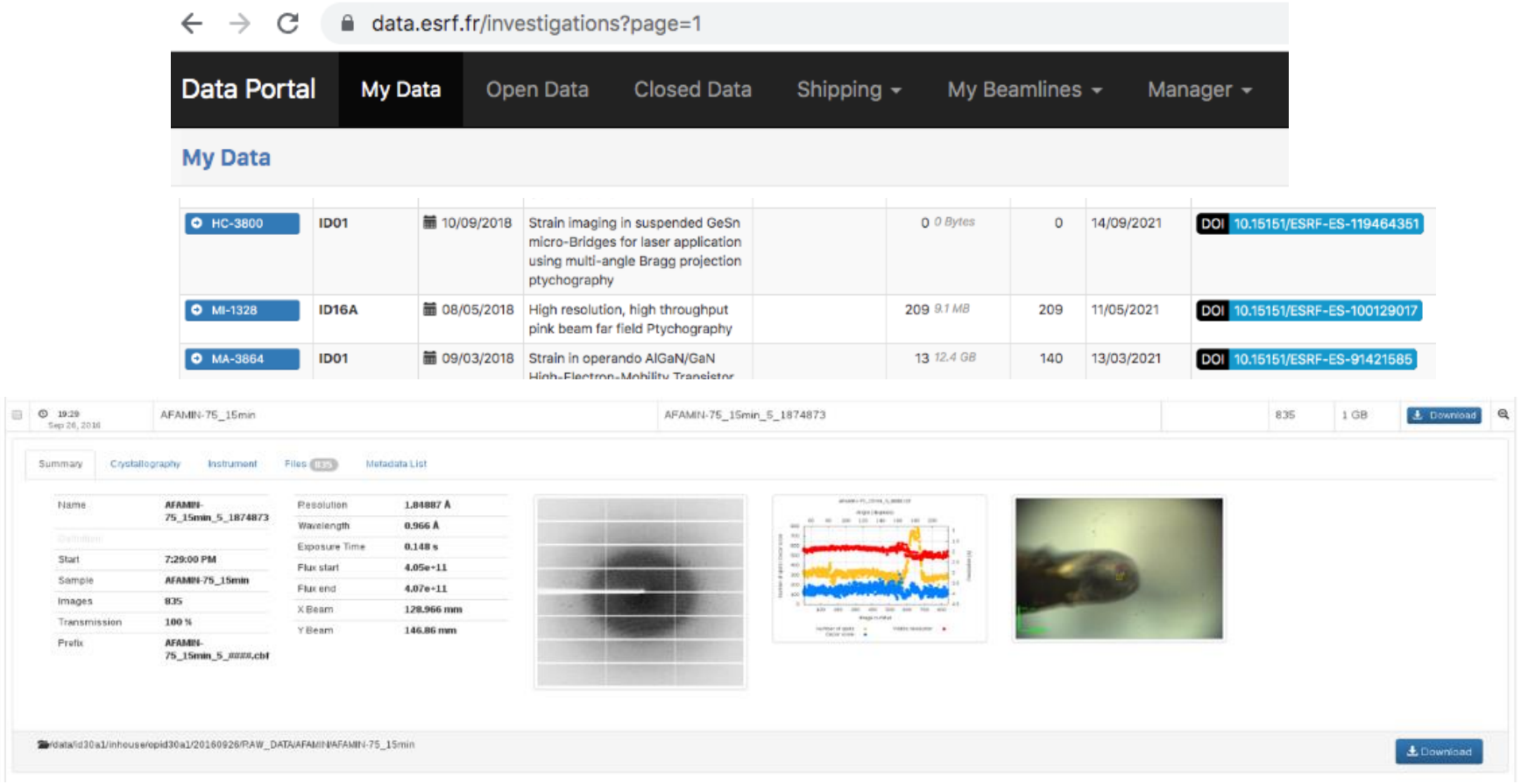
Data archiving

9. **Make clear how and when your data can be shared with scientists outside your group.**
10. **If your research involves sensitive data, explain any legal and ethical restrictions on data access and reuse.**
11. **Assign responsibility for long-term data curation to a suitable office.**

- Data need to be archived for long term future use
- You don't know when and how your data could turn out to be useful
- The meaning of long term depends on the data e.g. is 10 years enough?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

ESRF data portal - https://data.esrf.fr

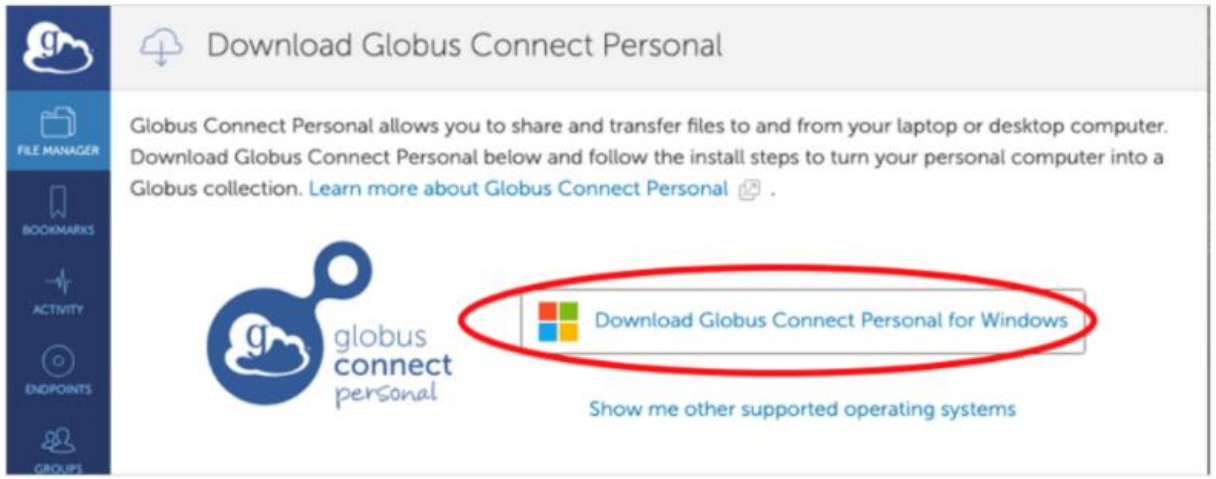


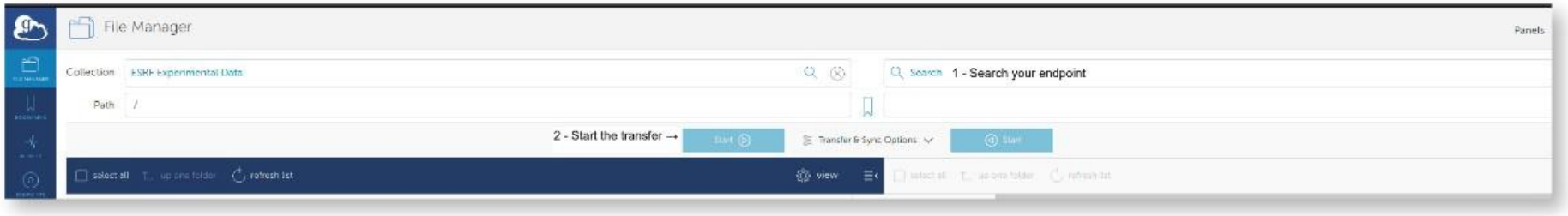
ICAT project collaboration https://github.com/icatproject

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Downloading large data: globus online

For users that want to download large volume of experimental data **(largest transfer so far 50TB)**





The service opened in fall 2021 for all users and all data
Data access is protected using Access Control Lists (ACLs) on the storage – users cannot see others data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Digital Object Identifier (DOI)

A DOI or Digital Object Identifier, is a string of numbers, letters and symbols used to permanently identify any object and link it to the web.

DOIs were originally used for publications and are now used for many things including movies, samples, instruments and scientific DATA.

- A DOI is one implementation of a PID (Persistent Identifier)
- A web address (url) is not a PID because it is not guaranteed
- Make sure the data you want to cite has a DOI
- Cite the instrument, samples etc. you used

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Journal require datasets accessible

More and more journals require datasets used in the publication to be cited and accessible. For example eLife, Nature, Plos, Science, …

- eLife – https://reviewer.elifesciences.org/author-guide/full

  All datasets used in a publication should be cited in the text and listed in the reference section and/or data availability statement. References for data sets and program code should include a persistent identifier, for example a Digital Object Identifier (DOI) or accession number.

  …

  Relevant software or source code should be deposited in an open software archive.

  …

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Example of article citing data

nature neuroscience

Explore content | Journal information | Publish with us | Subscribe

nature › nature neuroscience › technical reports › article

Technical Report | Published: 14 September 2020

Dense neuronal reconstruction through X-ray holographic nano-tomography

Aaron T. Kuan, Jasper S. Phelps, Logan A. Thomas, Tri M. Nguyen, Julie Han, Chiao-Lin Chen, Anthony V Azevedo, John C. Tuthill, Jan Funke, Peter Cloetens, Alexandra Pacureanu & Wei-Chung Allen Lee

Nature Neuroscience **23**, 1637–1643 (2020) | Cite this article

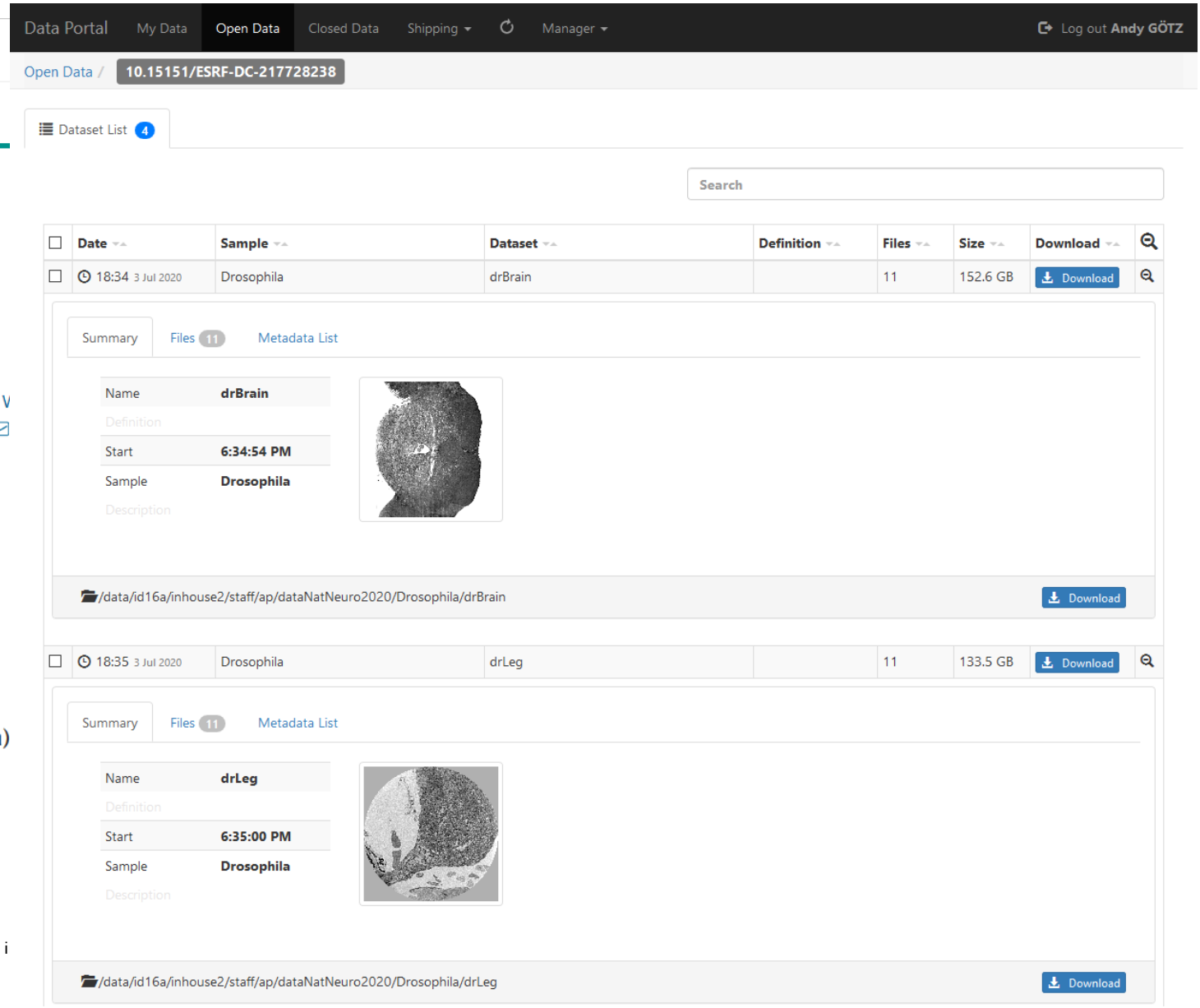
5492 Accesses | **8** Citations | **196** Altmetric | Metrics

https://doi.org/10.1038/s41593-020-0704-9

3. ESRF (https://data.esrf.fr/public/10.15151/ESRF-DC-217728238) (anonymous login)

DOI: doi.esrf.fr/10.15151/ESRF-DC-217728238

This project has received funding from the European Union's Horizon 2020 research and i



Data storage

4. **Decide what data and research materials require archiving and determine how much storage space you will need.**

- Data volumes are constantly increasing (up to Petabytes)
- You could be faced with more data than you can store locally
- Research facilities provide services to keep raw data at the facility
- Access to remote data is via remote data services (similar to cloud)
- Commercial cloud offer practically unlimited resources at a cost
- Data stored on commercial cloud disappear when you stop paying

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

File naming conventions

3. List the various types of data and research outputs that you expect to produce.

Adopt a directory and file naming convention which will allow you to know what the file contains.

- For example:

  Proposal/Beamline/Sample_name_Scan_type.ext

  MA1234/ID56/Gold_50_nm_ptycho_scan.h5

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Own your identity in the digital world

In a digital world you need to control your identity and not give it away to the corporate world to exploit. It is highly recommended to create your own identity using ORCID – a free non-commercial service

- Benefits of an ORCID identity:
  - You will be distinguished from every other researcher, even researchers who share your same name,
  - Your research outputs and activities will be correctly attributed to you,
  - Your contributions and affiliations will be reliably and easily connected to you,
  - You will save time when filling out forms, (leaving more time for research!),
  - You will enjoy improved discoverability and recognition,
  - You will be able to connect your record to a growing number of institutions, funders, and publishers,
  - Your ORCID record is yours, for free, forever.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

What are the advantages of producing FAIR Data?

- Better data and metadata means better science
- Saves you time and improves your results
- Allows you to use standard data services
  - Remote data analysis
  - Data archiving
  - DOI
- Publications with open data are cited more often
- You get more credit for your work
- Science is more reproducible and replicable

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Benefits of data sharing

Benefits of Data Sharing for Different Players in the Research Environment

Benefits for researchers:

- increases visibility of scholarly work;
- likely to increase citations rates, for example, open access journal articles are cited more;

(Continued)

(Continued)

- enables new collaborations;
- encourages scientific enquiry and debate;
- promotes innovation and potential new data uses;
- establishes links to next generation of researchers.

Benefits for research funders:

- promotes primary and secondary use of data;
- makes optimal use of publicly funded research;
- avoids duplication of data collection;
- maximizes return on investment.

Benefits for the scholarly community:

- maintains professional standards of open inquiry;
- maximizes transparency and accountability;
- promotes innovation through unanticipated and new uses of data;
- enables scrutiny of research findings;
- improves quality from verification, replication and trustworthiness;
- encourages the improvement and validation of research methods;
- provides resources for teaching and learning.

Benefits for research participants:

- allows maximum use of contributed information;
- minimizes data collection on difficult-to-reach or over-researched populations;
- allows participants' experiences to be understood as widely as ethically possible.

Benefits for the public:

- advances science to the benefit of society;
- adopts emerging norms such as open access publishing;
- to be, and appear to be, open and accountable;
- complies with openness laws and regulations.

"Managing and Sharing Research Data: A Guide to Good Practice" by Louise Corti etc al

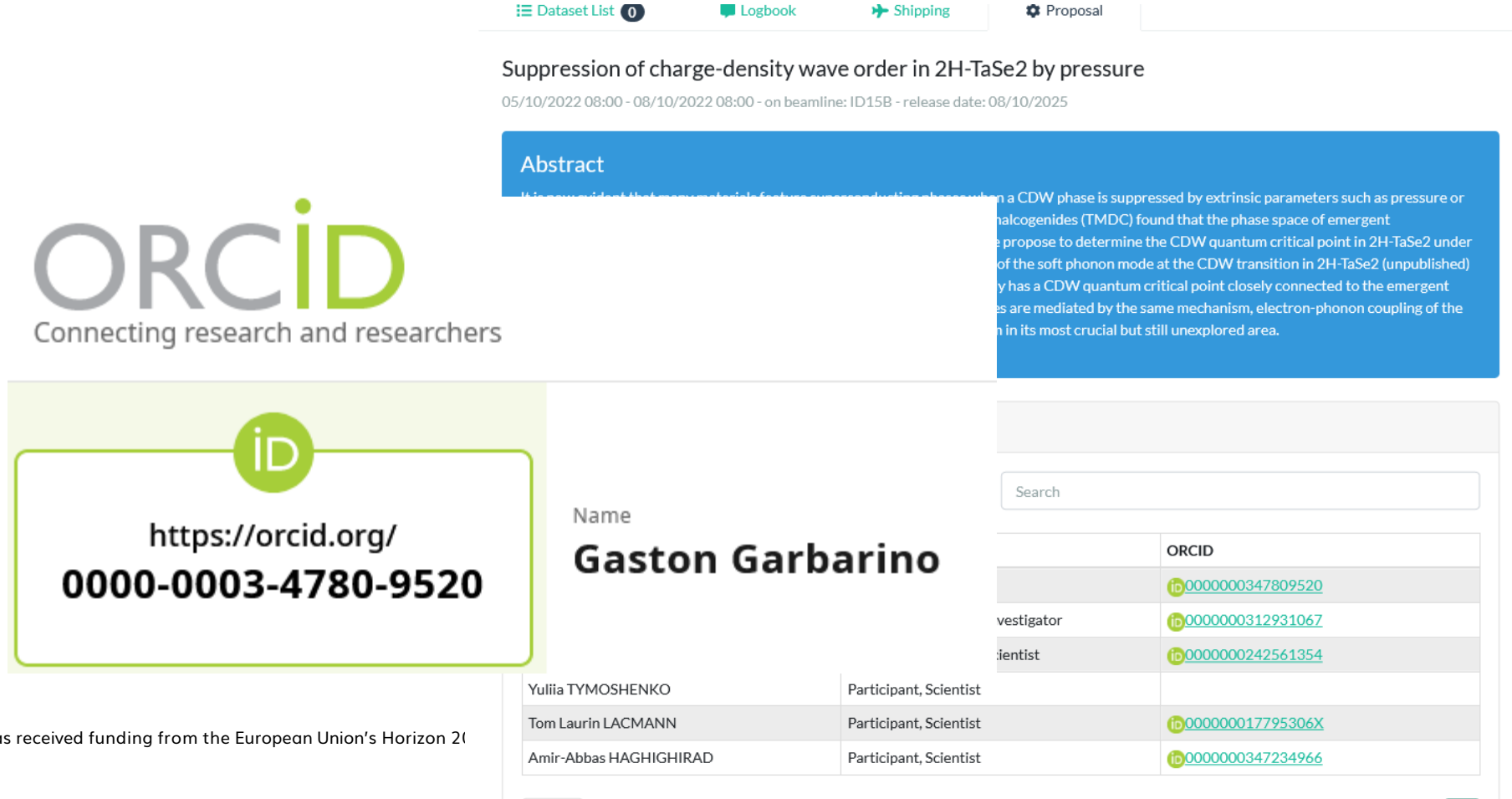
https://study.sagepub.com/corti2e

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Open identifier – ORCID.org

Achieving100% Open Identifiers:

1. *All scientists encouraged to create an ORCID*
2. *Encourage the use of ORCID for users for publications*



This project has received funding from the European Union's Horizon 2(

Open Training – https://pan-learning.eu

Catalogue | e-Learning | Events | About | Andy Goetz

PaNOSC summerschool FAIR session

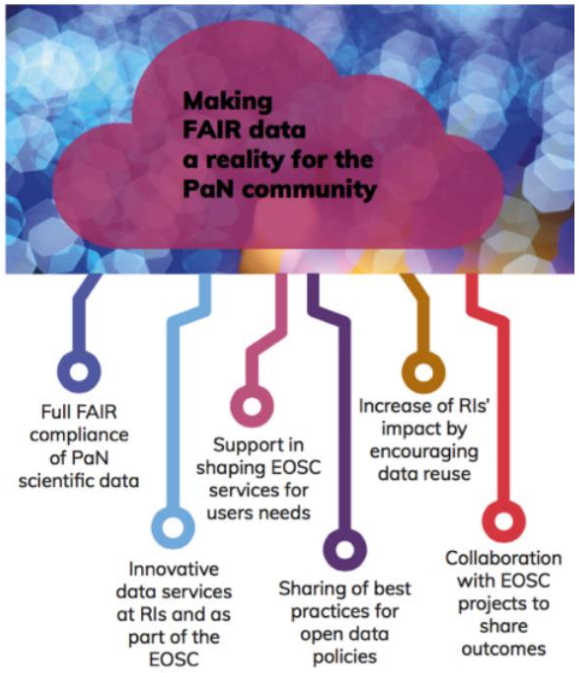
- Participants
- Grades
- Making FAIR data a reality
- Trust in Science
- What is Open Science
- What is EOSC
- What is PaNOSC + ExPaNDS
- Outcomes of PaNOSC + ExPaNDS
- What is Scientific Data and Metadata
- How to make data FAIR
- Your role as a researcher in Open Science
- Questions
- e-Learning
- Dashboard
- Calendar
- Content bank
- Quiz Taster

PaNOSC summerschool FAIR session

e-Learning | My courses | PaNOSC summerschool FAIR session

Turn editing on

Making FAIR data a reality



What are Science, Open Science, EOSC and PaNOSC

Please answer to the questions below to start the course.

Announcements

Catalogue | e-Learning | Events | About

PaNOSC summerschool FAIR session

- Participants
- Grades
- Making FAIR data a reality
- Trust in Science
- What is Open Science
- What is EOSC
- What is PaNOSC + ExPaNDS
- Outcomes of PaNOSC + ExPaNDS
- What is Scientific Data and Metadata
- How to make data FAIR
- Your role as a researcher in Open Science
- Questions
- e-Learning
- Dashboard
- Calendar
- Content bank
- Quiz Taster

Trust in Science

- Why trust
- Why trust science quiz

What is Open Science

- Open Science definition
- Pillars of Open Science
- Process of Openness in Research
- Open Science Schools of Thought
- Open Science Resources

What is EOSC

- The European Open Science Cloud
- EOSC Projects and Ecosystem

What is PaNOSC + ExPaNDS

- PaNOSC + ExPaNDS projects

Outcomes of PaNOSC + ExPaNDS

- PaNOSC + ExPaNDS FAIR Outcomes

What is Scientific Data and Metadata

- Scientific Data and Metadata

https://e-learning.pan-training.eu/moodle/course/view.php?id=11

Estimated carbon footprint of experiment

Calculated by Andy Götz

- Beamtime energy consumption = 2056 kg
- User Travel = 1170 kg
- Data stored on disk = 1.8 kg
- Data processing on site = 12.6 kg
- Cloud transfer = 2.3 kg

(CO2e per kwH in France = 75 g/kWh)

TOTAL = 3.253 tons !

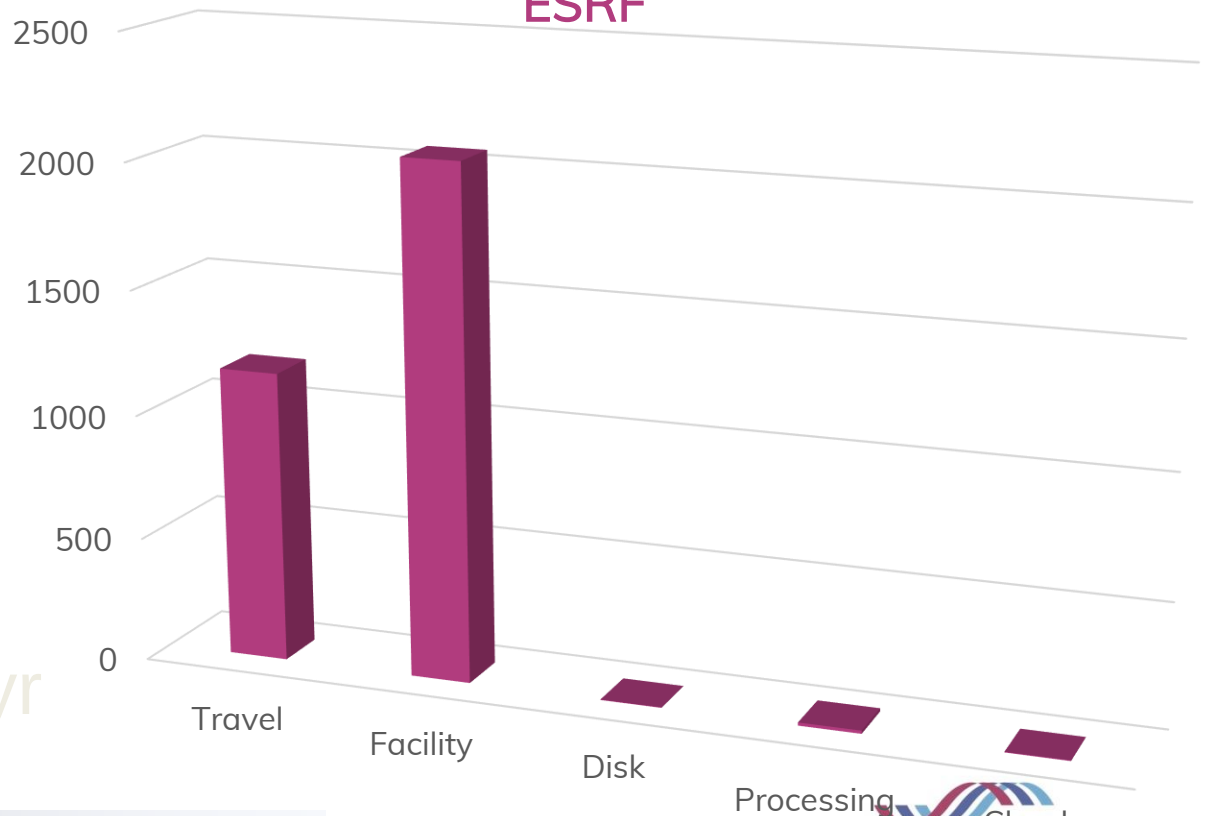
Sustainable Goal = 5 tons / human / yr

NEWS | 12 October 2022

Energy crisis squeezes science at CERN and other major facilities

LHC to end 2022 data-taking season two weeks early to save on electricity, among other measures.

Carbon footprint for 1 week experiment @ ESRF

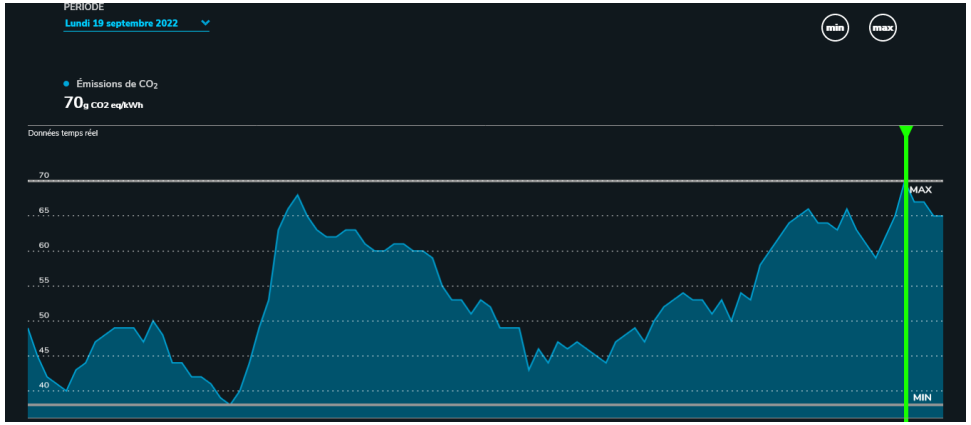


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Calculating the carbon footprint of data

- **User Travel -** 3 users fly from Copenhagen to ESRF (380+10 kg CO2e) = 3 x 390 kg
- **Beamtime energy consumption** – 1 week of beamtime (8MW/42) = 190 kWh
- **Data stored on disk** – 100 GB stored on disk (10W x 100 days)
- **Data processing on site** – 1 week of processing on 64 cores (1kW x 1 week)
- **Data transfer** – transfer 100 GB of data back to user (31 kWh)

CO2e per kwH in France (2022) = 75 g/kWh



This project has received funding

https://www.rte-france.com/eco2mix/les-emissions-de-co2-par-kwh-produit-en-france#

Carbon footprint of archiving data

- Data stored on tape for 10 years ~ 200 g * 35 = 7 kg

CO2e per kwH in France = **75 g/kWh**

ARCHIVING for 10 years ~ 7 kgs

i.e. 0.2% of the CO2e of the raw data!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data availability – the wrong + right way

Data availability

Data available on reasonable request to the authors.

Open Research

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6993871 , reference number 6993871.

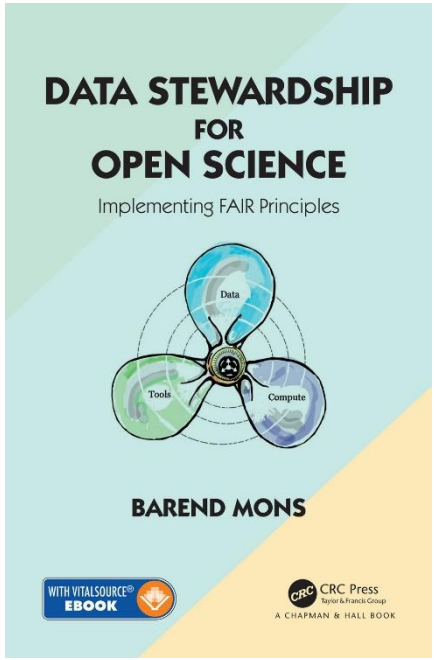
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc

Learning more about FAIR RDM for data managers

- RDMKit - https://rdmkit.elixir-europe.org/index.html
  - Provides a rich set of resources for all aspects of RDM mainly for researchers working in the Life Sciences but also for other Sciences. Very comprehensive overview, pragmatic approach, up-to-date. An excellent place to start and/or find information.

- Recommended reading:



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Tools to help you manage your research

A non-exhaustive list of tools to explore

- Open science framework – osf.io
- Protocols.io
- Fairsharing.org
- Jupyter.org notebooks

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Conclusion

Adopting best practices for Open Science and FAIR Data has many benefits especially helping MAKE BETTER SCIENCE

- Follow a **checklist** which covers the following topics:
  - Data Management Plan, Data Policy, Data Outputs, File types, File Formats, Software, Workflows, e-Logbooks, Data Storage, Data Archiving, Data DOI
  - Apply the **FAIR principles** – ask yourself if you or someone else will be able to use or understand your data
  - Make your Data **FAIR** – **release it and cite the data DOI**
- The **digital tools** exist for treating your data seriously
- There is a lot more to science than just text publications …

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

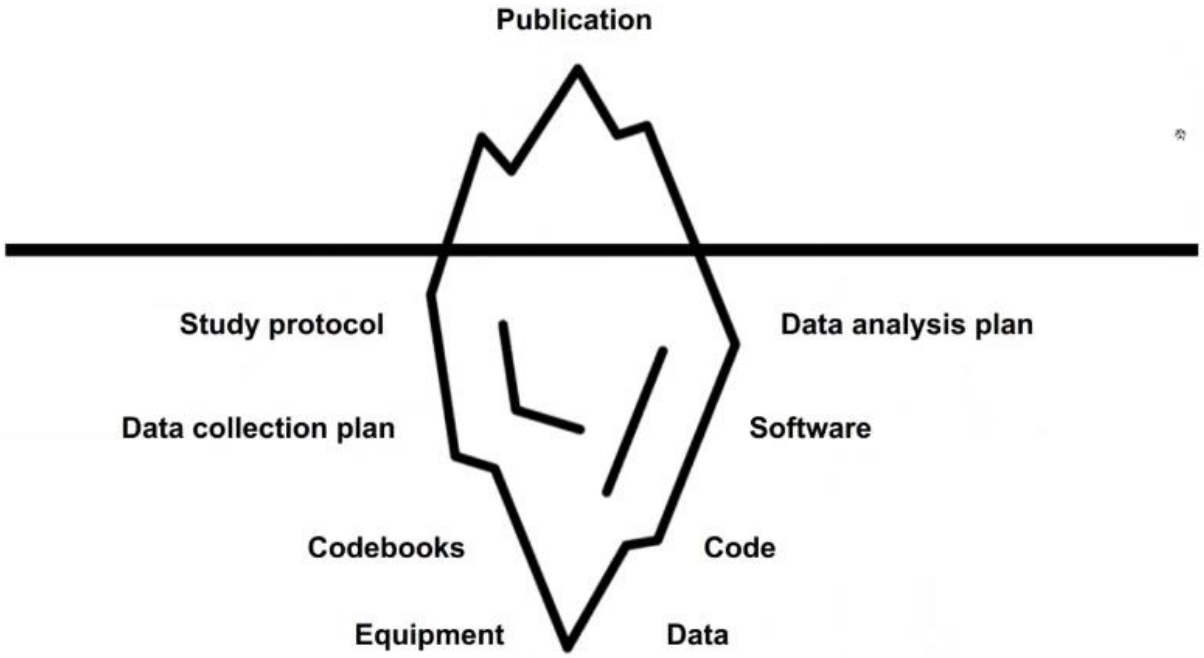
Acknowledgements

- RDMKit Elixir online guide
- University of Saskatchewan
  - https://library.usask.ca/studentlearning/workshops/grad-research.php#panel-section-3-ResearchDataManagementWhatYouNeedtoKnow
- **Nature** magazine, Scientific Data
- **PaNOSC, ExPaNDS, EOSC** H2020 projects
- **Wikipedia, Internet, ChatGPT**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Thank you

andy.gotz@esrf.fr



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data management made simple

Quirin Schiermeier in Nature (2018)

CAREER FEATURE | 13 March 2018

Data management made simple

Keeping your research data freely available is crucial for open science – and your funding could depend on it.

https://doi.org/10.1038/d41586-018-03071-1

1. Check the research-data requirements of your funding agency and field of research.
2. Go online for help in developing a data-management plan. A useful guide outlining UK funder expectations can be found at go.nature.com/2tnohla.
3. List the various types of data and research outputs that you expect to produce.
4. Decide what data and research materials require archiving and determine how much storage space you will need.
5. Define appropriate data file formats (see https://fairsharing.org/ for formats).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Data management made simple

Quirin Schiermeier in Nature (2018)


CAREER FEATURE | 13 March 2018

Data management made simple

Keeping your research data freely available is crucial for open science – and your funding could depend on it.

https://doi.org/10.1038/d41586-018-03071-1

6. Look for data repositories used by your research community or your host institution (see www.re3data.org for examples).
7. Check what data format and structure the chosen archive might request.
8. Provide metadata that allows others to understand, cite and reuse your data files.
9. Make clear how and when your data can be shared with scientists outside your group.
10. If your research involves sensitive data, explain any legal and ethical restrictions on data access and reuse.
11. Assign responsibility for long-term data curation to a suitable office.
12. Revisit your plan frequently and update it if necessary.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852