Data Policy at Large Research Infrastructure. Data Retention challenge.

WORKSHOP ON *WHEN SHOULD SMALL MOLECULE CRYSTALLOGRAPHERS PUBLISH RAW DIFFRACTION DATA?*

11-12 August 2021

Outline

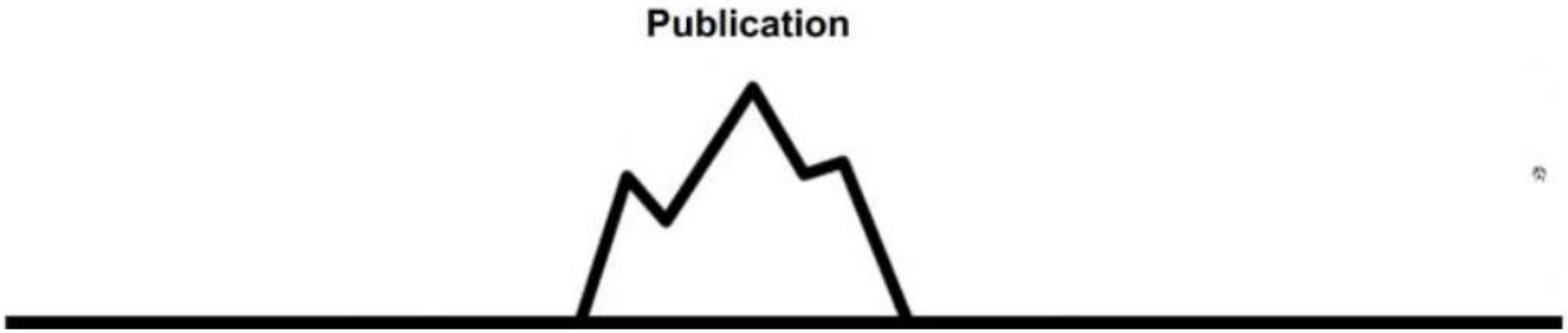
1. Why FAIR and Open Access
2. FAIR data policy, DMP and Open Access
3. What FAIR brings/what data to publish/how to publish
4. ELI/PaNOSC use-cases
5. Conclusions

1. HTTPS://EN.WIKIPEDIA.ORG/WIKI/REPLICATION_CRISIS

Why FAIR?
Why OPEN ?

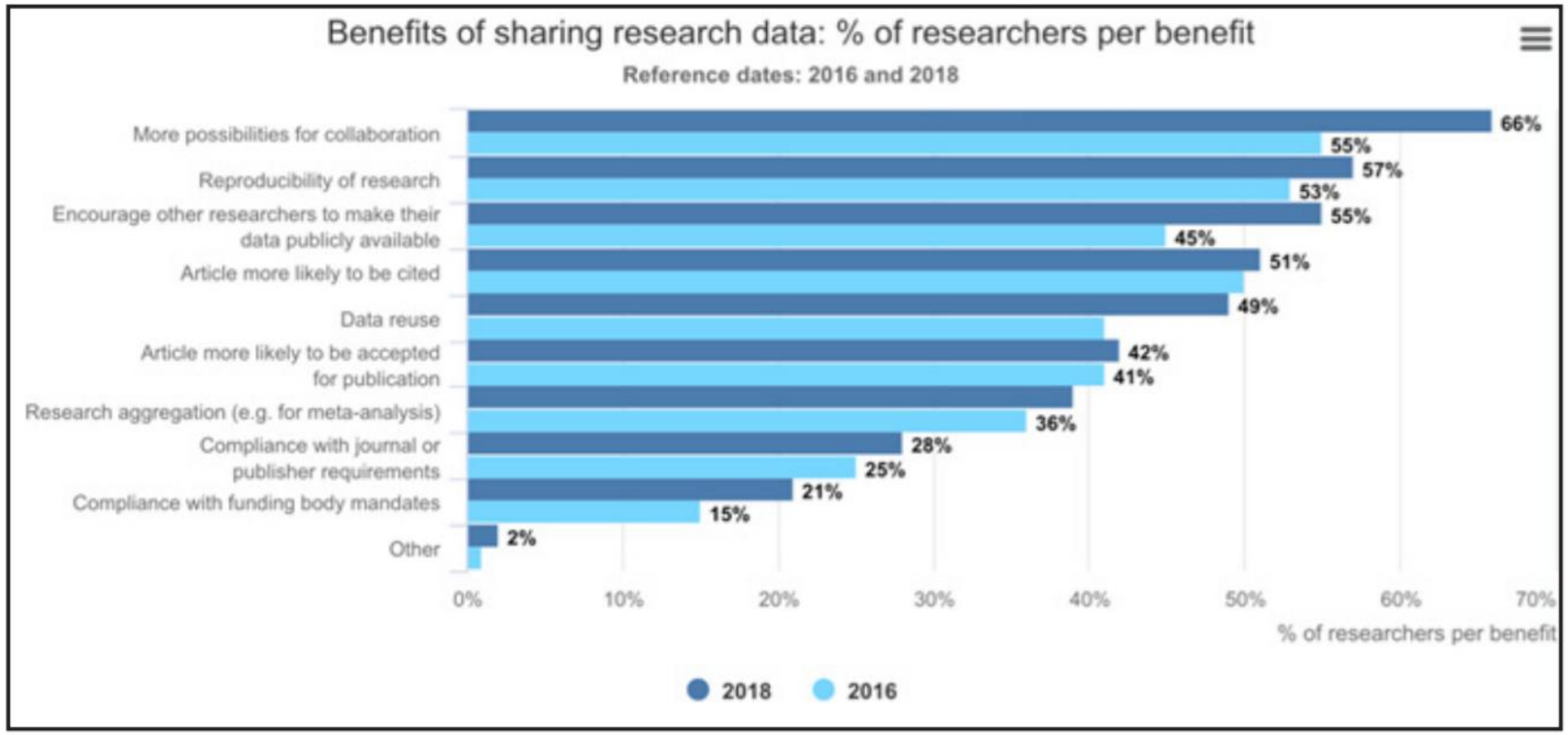
Science produces much more than Publications

Science produces Publications



“Ideas, theories, and their supporting intellectual frameworks should cc
a common good of all humanity, freely shared for our mutual enjoyment and
benefit.”[1]

FAIR and Open Science Impact!



More details: https://sciencebusiness.net/report/why-open-science-future-and-how-make-it-happen

Scientific impact and cost effective:

The European Union commissioned a study of the cost to the research community from not having FAIR data[1] .

The study estimates a cost impact to the entire European research community of over €10bn per year. This cost impact provides a market valuation of research data, that without FAIR, is not curated and has degraded quality.

1. HTTPS://OP.EUROPA.EU/S/PEVT

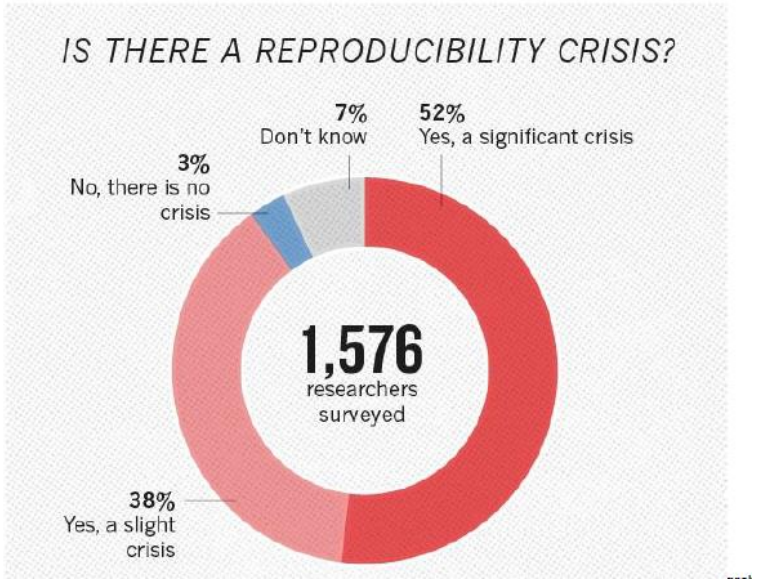
Reproducibility crisis???

No raw data, no science: another possible source of the reproducibility crisis[1]

Computational physics – only 12 results reproduced out of over 300? [2]

Is there a Reproducibility Crisis? What are the causes? How can it be addressed?

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



What are the factors (https://www.nature.com/articles/d42473-019-00004-y): (ex comes from biomedical but conclusions are accurate!)

- *A lack of access to methodological details,* ***raw data****, and research materials*
- *Use of* ***misidentified****, cross-contaminated, or over-passaged cell lines and microorganisms.*
- ***Inability to manage complex datasets***
- ***Poor research practices and experimental design***
- *A competitive culture that rewards novel findings and undervalues* ***negative results (But are there really NEGATIVE Results or is it just a different perspective?)***

HTTPS://EN.WIKIPEDIA.ORG/WIKI/REPLICATION_CRISIS
HTTPS://PHYS.ORG/NEWS/2017-03-SCIENCE-CRISIS.HTML

It all leads to (RAW) Data!

1. https://pubmed.ncbi.nlm.nih.gov/32079532/
2. https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0069

FAIR principles, present in the Data Policy is becoming THE community standard!

PaN Partners are also approaching the Open Access challenge!

ISIS neutron and Muon facility	no guarantee - long term archive	3 years	https://tinyurl.com/f3zhnpw3
ESRF	5 years minimum, 10 years expected	3 years	https://tinyurl.com/3rpe9vk6
ILL	5 years minimum, 10 years expected	5 years	https://tinyurl.com/2afuk755
Sol eil	5-10 years	3 years	https://tinyurl.com/48vb9f73
DESY	Not specifically defined	Not Specifically defined	https://tinyurl.com/hrr4nzpb
FRMII	10 years	Not Specifically defined	https://tinyurl.com/tdkn67y9
HZDR	10 years	5 years	https://tinyurl.com/4brvdtuv
HZB	10 years	5 years	https://tinyurl.com/n62tnv62

PaNOSC - of the surveyed facilities with a specific policy for scientific data, 15 facilities have specific terms for open access to data.

There's still some work to be done:

- The existence of a given facility's data policy does not necessarily guarantee that the policy terms are fully implemented.
- This requires considerable resources for both staff and in capital investment, which are tensioned against other aspects of a facility's operation.

What are the challenges:

- What's the cost, how you make the
- Know your DATA becomes the most critical aspect

FAIR Data Challenges

1. FAIRification of existing data and upgrading systems for facilities already operating

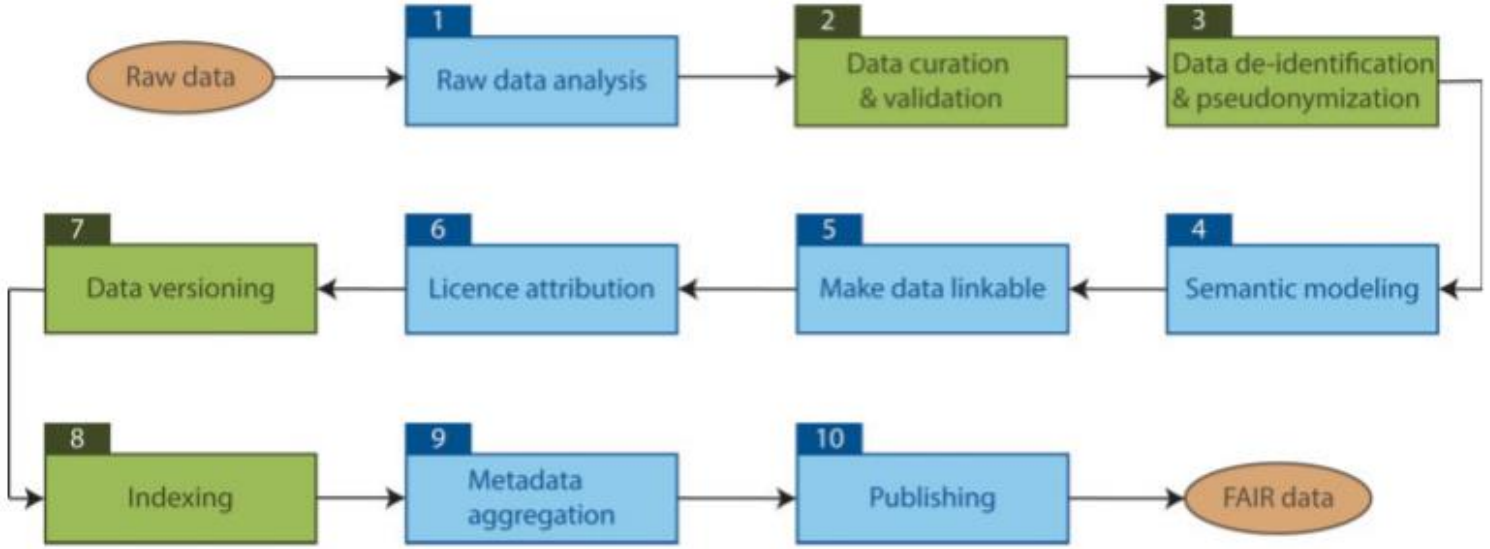


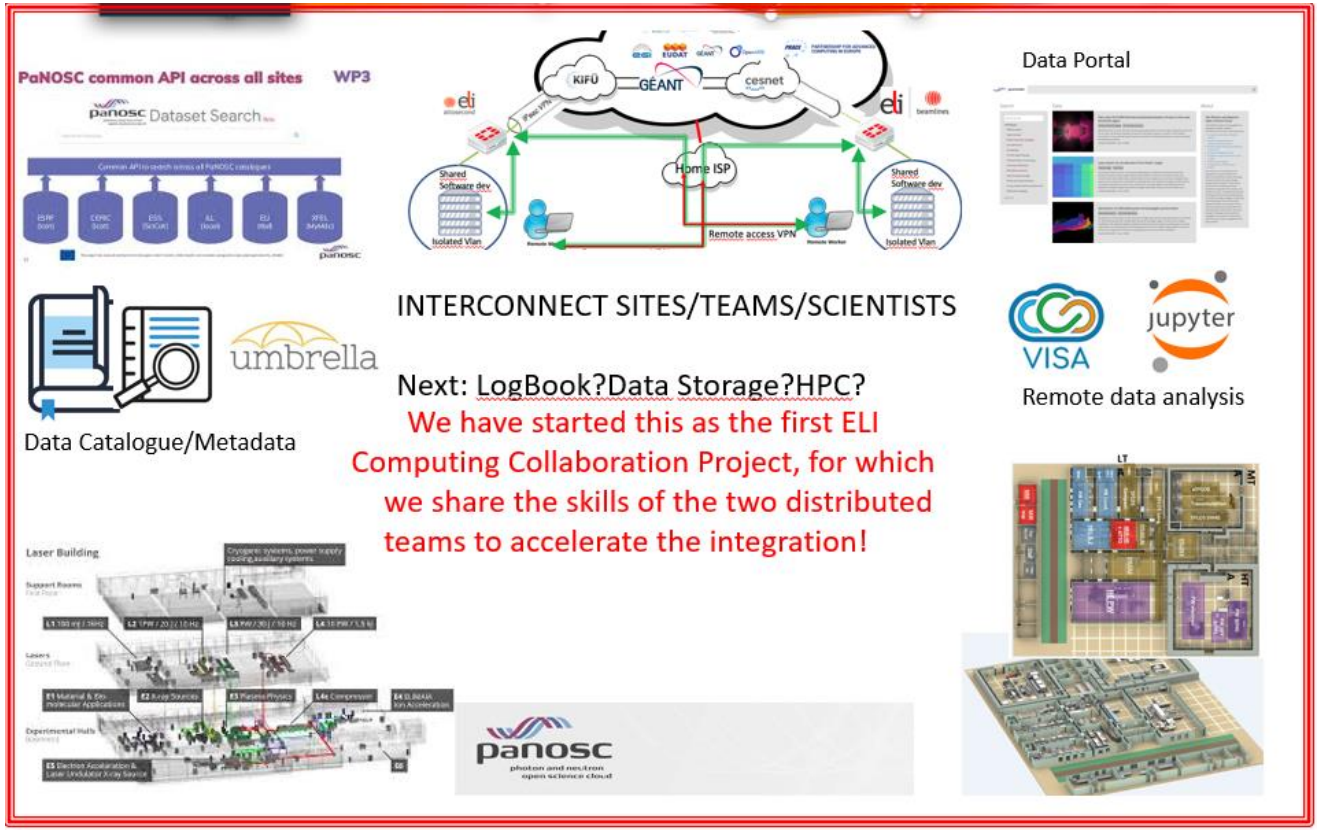
Fig. 1 The FAIRification workflow for health data. The blue boxes come from the GO FAIR (an initiative implementing FAIR data principles) process, while the green boxes are newly introduced steps to meet the specific challenges of health data. FAIR, findability, accessibility, interoperability, and reusability.

Costs challenges:

- There are no clear cost drivers – PaNOSC preliminary analysis = different IT strategies => different cost impact
- Adds extra IT&C requirements
- Data Acquisition/Data Ingestion pipelines will need upgrades
- **CS upgrades are surely needed.**

http://dx.doi.org/10.1055/s-0040-1713684

FAIR Data Challenges



2. Some facilities are using the timing advantage, aiming to **be FAIR by Design.**

FAIR is easier for new RIs:

- Using FAIR as standard in designing and implementing Scientific Data Management Tools
- Increases the load on CS and DaQ groups.

Challenges:

- Computing models are not yet well-known
- New users, new physics, new requirements
- Raw data is challenging. What is raw data ?

How it is for ELI?

We aim to start with FAIR data for FAIR experiments!
- Data policy framework[1] – PaNOSC
- PaNOSC guidelines on Best Practices implementing a research data poliy[2]

Article 13 Data Policy

13(1) **'Data' refers to all information collected by USERS and the staff while performing scientific experiments under the ACCESS FOR USERS Policy and performing operations of the ELI FACILITIES.**

13(2) **Open Access to FAIR data sets and metadata stored in Open Access repositories shall be favored** for data collected as a result of the use of the ELI FACILITIES and, to the extent possible in case of software and computer programs created by the ELI ERIC and the ELI FACILITIES; open source principles shall be considered.

1. https://www.panosc.eu/data/panosc-data-policy-framework/
2. https://www.panosc.eu/news/guidelines-for-implementing-a-research-data-policy-released/

Research Data Management Plans During The Project Life Cycle, **describes Embargo and impacts the Data Retention!**

Why should you create a DMP?

- Check you have the necessary equipment and support in place
- Decreasing the risk of duplication, data loss or data security breach
- Ensuring that the data are complete, accurate and reliable
- Saving time and energy (e.g., when searching the data, writing up papers, etc.)
- Some funders require a DMP (e.g., Horizon 2020)

Most importantly, the DMP helps you prepare and understand the Data Challenges!

- Do you have enough storage, or will you need to include charges for additional services?
- Who will be responsible for backup and recovery?
- What are the risks to data security and how will these be managed?
- How will you ensure that collaborators can access your data securely?

https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

ELI USE CASE!
FAIR is team effort!

Raw = huge disk space
Curated=users' input
Retention depends on raw/curated definition
storage is a considerable cost driver!

Data Analysis tools/methods are needed!

Facility Timing Systems

White Rabbit was built for synchrotrons, not lasers (RF-driven, not event-driven)

- Relative Distribution: Our current solution is OK for 1 source, but not sustainable for 2 laser-experiments.
- Absolute Distribution: Finally technically solved (Rollout in late 2020/early 2021)

Integration of detectors & digitizers

Obviously requested
Next system: Andor

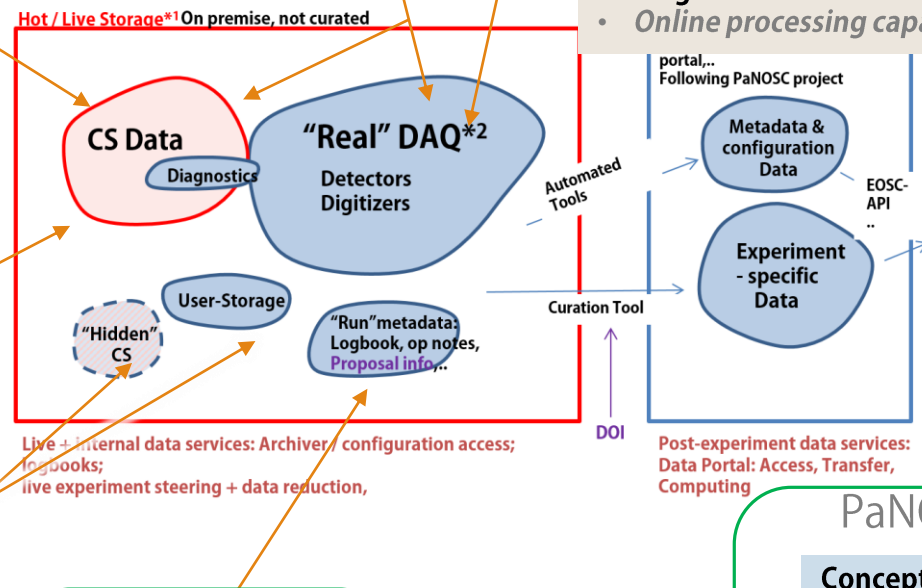
Storage + Infrastructure Development

- Capacity
- Peak Datarate
- Single Source Datarate
- *Online processing capacity*

Lead times for DAQ infrastructure is very long compared to commissioning timelines.
Balanced strategy between integration, data reduction, capacity growth is difficult to find.
Cost of systems goes down 15-30%/year..
Too early = outdated / limited expensive systems; too late = can't do science.

Homogenization Lifecycle Management

- We don't use the same archivers everywhere. Developing multiple high-level tools would become an unmanageable effort.
- A lot of data (beam images!) quickly loses its relevancy. If we don't manage that in 2021, our storage needs will explode.



Integration + Standardization Metadata Management

Process + Conceptual work

Many facilities provide a "user drive" for every visiting scientists. This is a useful prototyping project that helps us develop concepts for federated access, and it's a decent, although limited way to deal with unintegrated subsystems for a while.

Stress tests show that we could produce over 150 TB/week

Internal Data Services / Aggregation:
Logbook + Shot Report
MVP Metadata database

We're developing a framework that links all of our data to "outside"/high level services. Some of them are of operational nature (prototypes: servicestatus, "playground", configuration database), some are user-facing:

- Logbook, Shot report

MVP Data Transfer

We're already producing so much data that there's no choice about it.

PaNOSC

Conceptual work on

- Data Policy
- File Formats
- Metadata formats
- Unique Identifiers
- Nomenclature

A lot of this is in-sync with major other facilities (PaNOSC, ExPANDS) and most on ELI-ERIC level

- Data policy significantly affects obligations ("FAIR")
- Metadata + UIDs + formats + .. affect current service development

The Data Policy shows what's available, first user requirements are collected in a DMP!

What FAIR brings, what to publish?

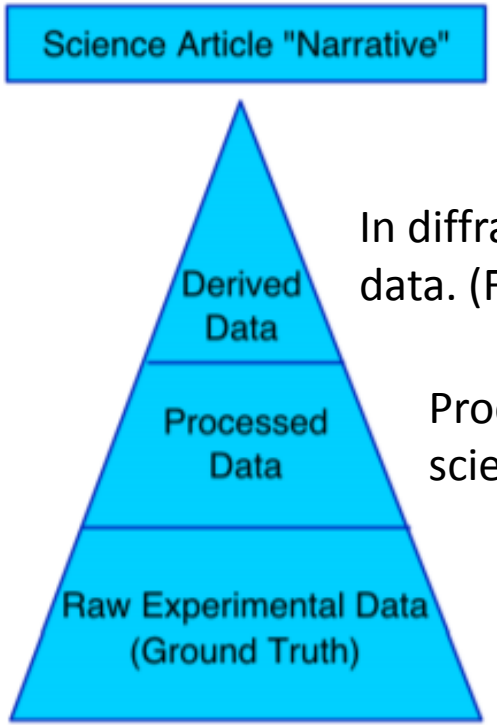


Fig1 – Data Pyramid

In diffraction applications, models such as molecular structures are then derived from the processed data. (For HEP = users' data @CERN, sometimes referred to as data analysis results)

Processed (also sometimes termed "reduced") to correct for artefacts and convert to scientifically meaningful units

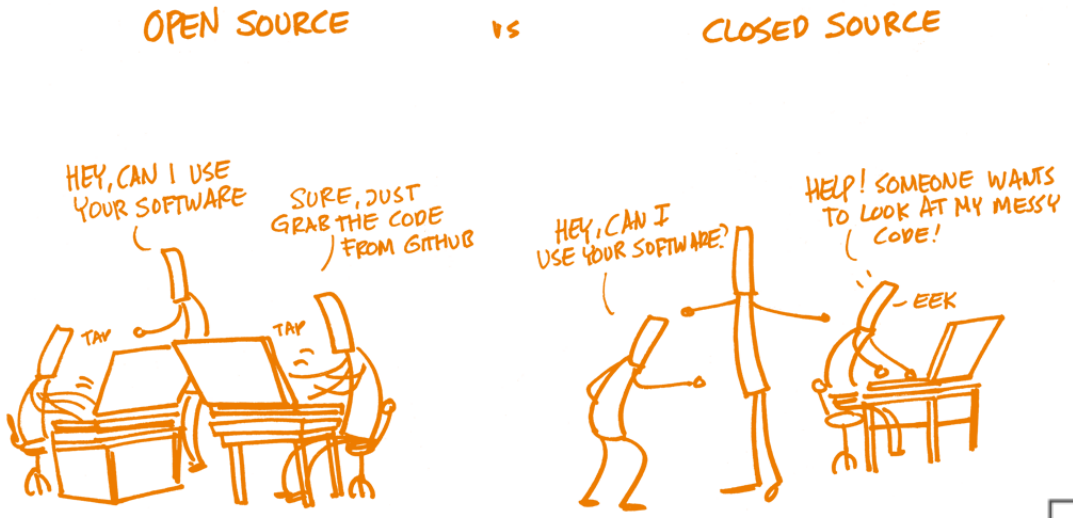
"Raw" experimental data are read directly from sensors or detector, with no or little conversion.

Union of Crystallography described the importance of quality *" the essential component of openness is that the data supporting any scientific assertion should be*

- ***complete*** *(i.e. all data collected for a particular purpose should be available for subsequent re-use); and*
- ***precise*** *(the meaning of each datum is fully defined, processing parameters are fully specified and quantified, statistical uncertainties evaluated and declared)."*

Fig1 Zenodo. https://doi.org/10.5281/zenodo.5155882 - Gotz, Andy, Helliwell, John, Richter, Tobias, & Taylor, Jonathan. (2021). The vital role of primary experimental data for ensuring trust in (Photon & Neutron) science.

Sharing is more than Data! Software tools!



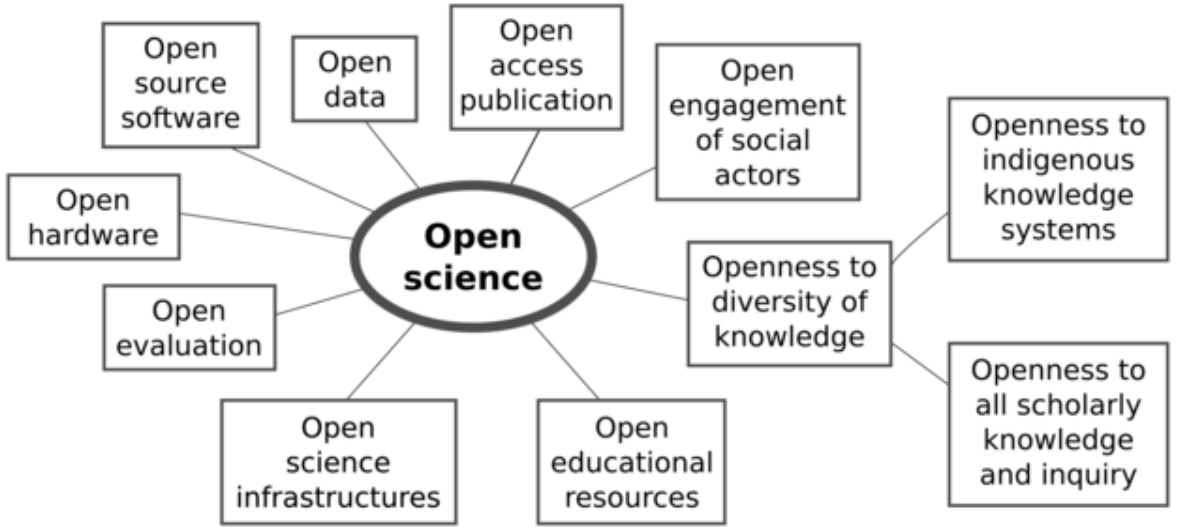
https://open-science-training-handbook.gitbook.io/book/open-science-basics/open-research-software-and-open-source

Users develop their own data analysis tools:

- Some are available – we need to collect (DMP?)
- Some not open – could be translated/improved
- Again, it all depends on the scientists!

The six principles of open science are

- Open methodology
- Open source
- Open data
- Open access
- Open peer review
- Open educational resources



By Robbielan Morrison - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=100144897

https://www.wiley.com/network/researchers/licensing-and-open-access/7-drivers-of-open-science
https://en.wikipedia.org/wiki/Open_science

Open Science – driven by the users!

“The goal is to turn data into information, and information into insight.” – Carly Fiorina (former CEO HP)

Carly Strasser, who oversees the foundation's Data-Driven Discovery Initiative. “Open science, data sharing, software sharing is the future of science,” she says. “It's only going to get more difficult to engage in science without being open.”

https://www.nature.com/articles/nj7584-117a

Scientists are stakeholders but, at the s
time, they are the **KEY CONTRIBUTOR!**

Figure 7: Have you done any of the follo

Figure 42: Journals by code sharing modality

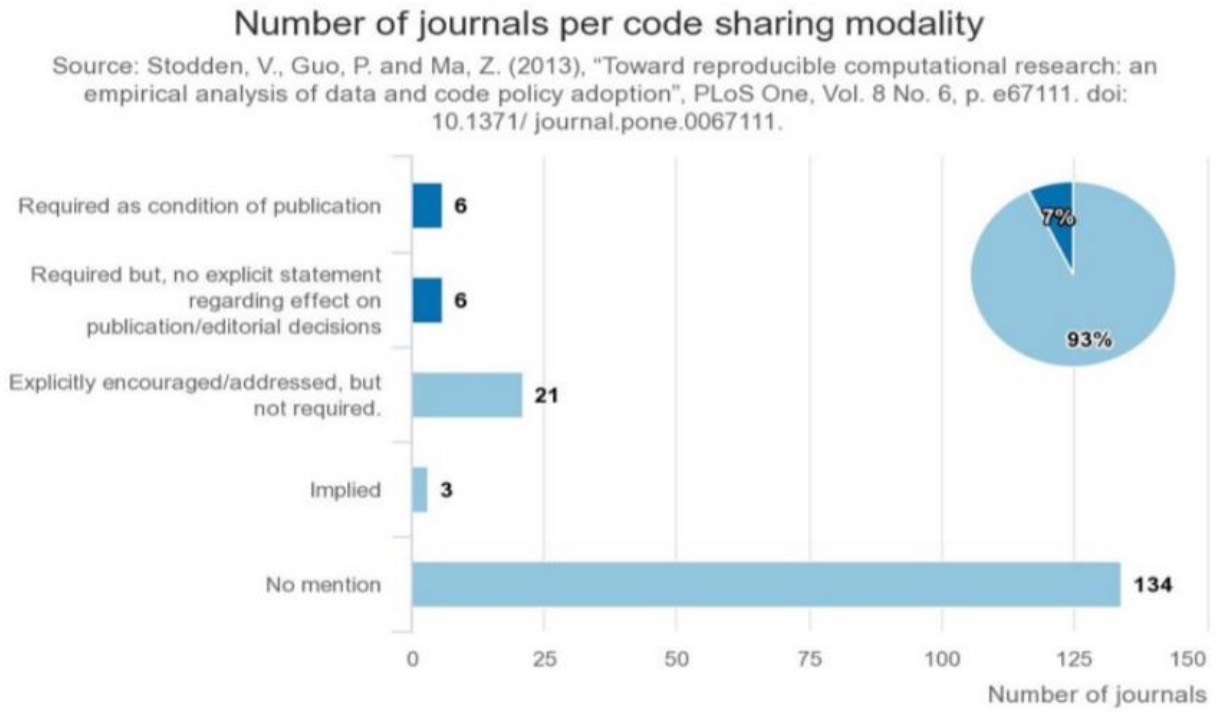
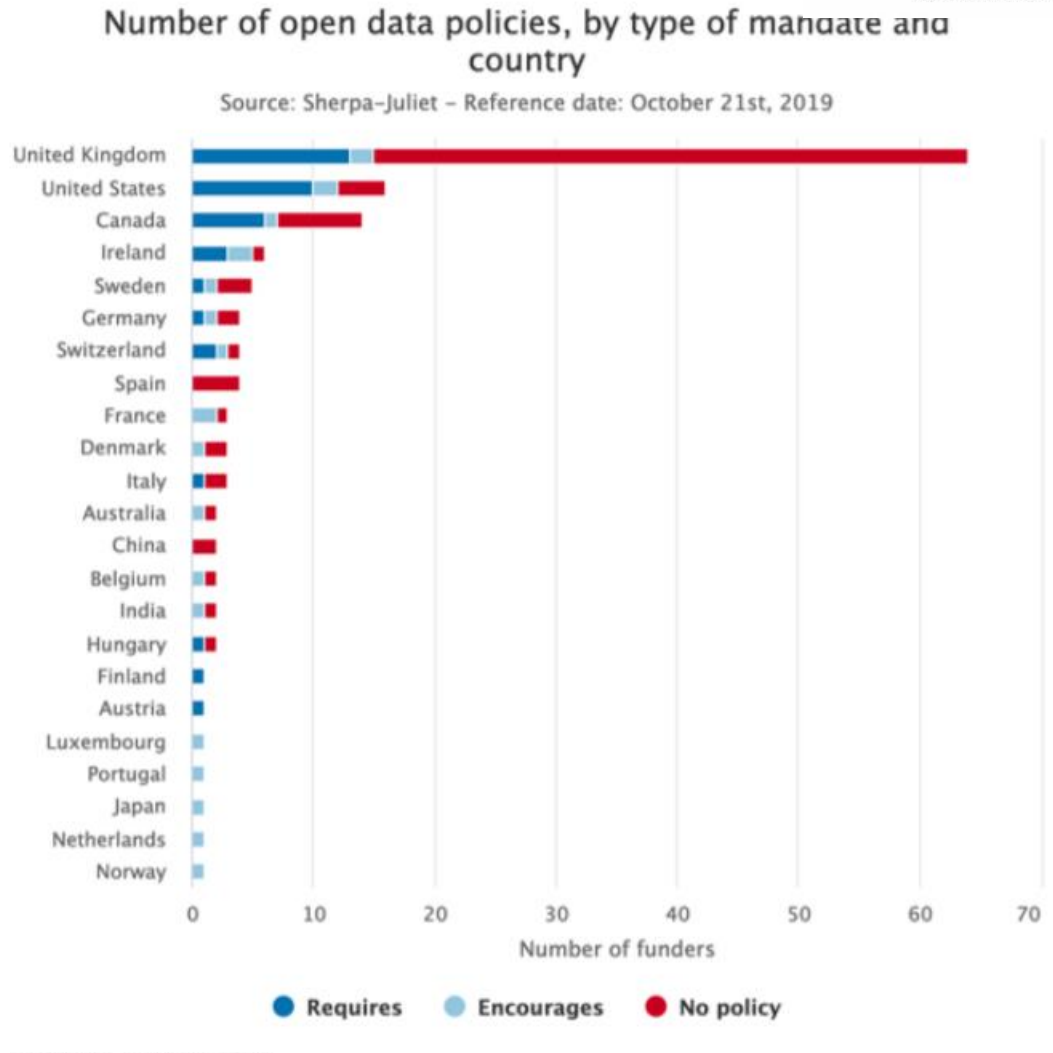


Figure 40: Open data policies by country

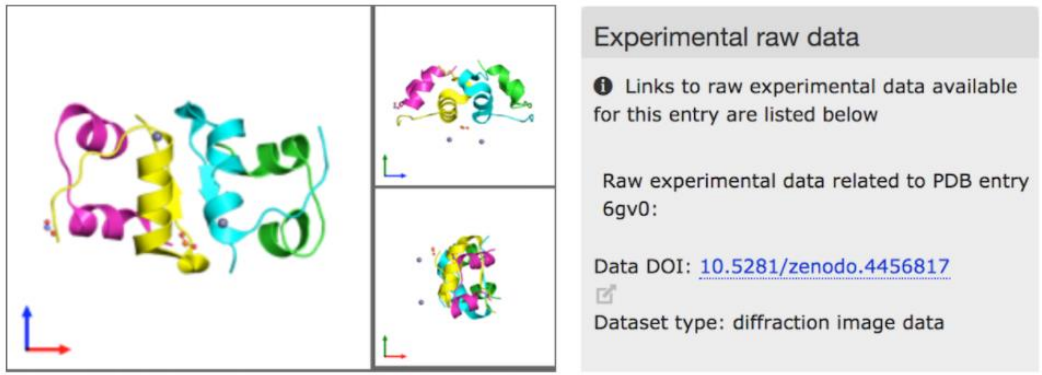


https://ec.europa.eu/info/sites/default/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_open_science_monitor_final-report.pdf

What FAIR and Open Data requires? …..Champions and use-cases!

This use case is to link raw data in the ESRF data repository to the PDBe entries as shown in the example here.

PDB 6gv0 coloured by chain and viewed from the front



Generalisation of the use case

All PaNOSC and ExPaNDS partners could add the adapter for linking PDB entries to raw data. Thereby making PDB entries FAIRer.

Resources

- ESRF data repository + PDBe adapter
- pDB 6gv0 structure summary

Description of needs

Modify PaNOSC data repositories to add support for linking PDBe entries to raw data.

Use case action flow

1.ESRF data repository – add javascript adapter for PDBe
2.PDBe repository – test adapter
3.Scientists – add DOI to raw data to PDBe entry

Impacts from the implementation

Make Protein Data Bank entries FAIR by providing access to raw data. This will enable results to be verified and structures to be refined with new software. This ICUR CommDat committee has been a strong advocate of making raw data accessible for crystallography [1]

[1] https://doi.org/10.1107/S2052252519005918

https://www.panosc.eu/use-cases/panosc-use-case-10-linking-raw-data-to-the-protein-data-bank-in-europe-pdbe/

Summary and conclusions.

PaNOSC

1. Data Policy updates ads an interesting change – the extension to processed data for publications!
2. Fairification requires time and planning(new challenges/new requirements/new cost drivers). Easier for “younger” facilities.

What is needed to approach FAIR/Open:

- RAW **Data,** Curated-processed **Data,** Standard **Data formats,**…
- It all starts with **Data and Use Cases!**

Science produces much more than Publications