The vital role of primary experimental data for ensuring trust in (Photon & Neutron) science

Paper written for the PaNOSC[1] project by
Andy Götz[2], John R Helliwell[3], Tobias Stefan Richter[4], Jonathan William Taylor[4]

1. Challenges for the Trust in Science

The foundation of experimental science is data and the progress of science relies upon a study being reproducible by others' analysis of the same data. Replicability, on the other hand, comes from independent groups performing their own experiments on the same phenomenon.
Reproducibility is not guaranteed in science and recently reproducibility, let alone replicability, has been called into question. A recent Nature group survey [1] indicated that a majority (70%) of respondents had failed to reproduce published results and many (>50%) could not reproduce their own experimental results. Whilst a single survey cannot be taken as indication of an endemic issue, improvements in key areas such as data management can ameliorate trust in science.

Science at photon and neutron facilities requires large investments for the sources, experimental stations, etc as well as significant running cost. Upgrades and new sources are necessarily brighter, this together with improvement in detector technology produces rapid increases in data rates. Dealing with the associated data volumes is becoming a burden for researcher and budgets. Traditionally cost of data storage was a small fraction of the expenditure of a facility. On the latest upgrades to the facilities Physics Today reported recently [2] that they will present a raw data deluge of such magnitude it was stated quite simply that the objective of full data archiving might be the ideal but it is impractical.
For example the estimated raw data volumes for data taken with the new ESRF

[1] https://panosc.eu
[2] ESRF, [3]Manchester University, [4]ESS

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

source (EBS) range from 10 to 100 Petabytes in 2021, and 50 to 500 Petabytes in 2025. The large range is dependent on data compression rates which are achieved. These can vary between 10 and 1000 depending on the techniques and data. Despite the very impressive upper limit of 0.5 Zetabyte (uncompressed) per year , tape storage (the main medium for long term storage) continues to advance and promises a factor of at least 10 increase in tape density over the next decade (see picture below). ESRF has 150 Petabytes of tape storage currently. Therefore technology is not the limitation in providing primary experimental data as the ground truth for science.

This accessibility to the primary experimental data cuts to the core of the validity of science. Curating raw data and making that data openly accessible for scrutiny or reuse is as important to the validity of science as the process of peer review. The importance of data is described in the European Code of Conduct for Research Integrity [3] which provides guidelines for essential data management considerations, such as ensuring research data are FAIR (Findable, Accessible, Interoperable and Reusable).

The move towards open data has many drivers from the research community, funders and governments. Moral obligations based on use of public funds for research and verifiable research are key considerations. Many journals allow (indeed require) supporting data as part of the submission process. For other researchers to have a real chance to reuse open data that data needs to be FAIR. The level of reuse of open data is often a matter for debate. Irrespective to this there are compelling arguments based on scientific best practice for a move towards open science, open data and FAIR. This necessitates a change to far better metadata collection and a greater emphasis on data curation and persistence.

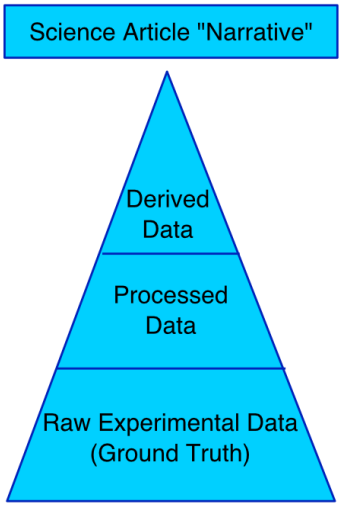
The European Union commissioned a study of the cost to the research community from not having FAIR data [4]. The study estimates a cost impact to the entire European research community of over €10bn per year. This cost impact provides a market valuation of research data, that without FAIR, is not curated and has degraded quality. The cost drivers are largely from the work required to ensure scientific data validity and scientific quality. Often infrastructure costs required for data curation seem prohibitive and arguably for research infrastructures are often not the highest priority. The integral cost of not implementing FAIR to the entire research community is high and far exceeds the investment costs of implementation which can be estimated by assessing the funding for the European Open Science

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

Cloud during the Horizon 2020 funding period, of ~600 €M.

In this article we describe the case for best practice of research data management in the photon and neutron community. Specifically, we discuss the importance of primary research data curation as seen from the perspective of the research publication, as well as assess the research data management implementation across the domain landscape. We describe some specific examples with an analysis of the costs vs benefits.

2. The Data Pyramid



In the photon and neutron community classes of data can be organised in tiers in a so called 'Data Pyramid' (Figure 1) or hierarchy to compare their relative impact to scientific validity and reproducibility. "Raw" experimental data are read directly from sensors or detector, with no or little conversion. In some cases, data labelled "raw" can already undergo well established and tested correction for issues like non uniformity or spatial distortion from within the detector readout pipeline. The raw data are then processed (also sometimes termed "reduced") to correct for artefacts and convert to scientifically meaningful units. In most areas of science there are multiple workflow possibilities and often a choice of software algorithms. In diffraction applications, models such as molecular structures are then derived from the processed data. While the volume of the dataset decreases in most cases as they are subjected to consecutive steps of processing – hence the "Data Pyramid" - the number of human decisions made in this work flow steadily increases the subjective nature of any analysis.

Recording and reporting of the workflows applied to traverse the pyramid is important, indeed vital, to the reader, user or reviewer of a research publication. Of course scientific understanding comes mainly from the final analysis, often a fitted model to processed data. But trust comes from the reproducibility of the data chain afforded only by curating the raw data. In a scientific study, the article is a narrative

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

written by the authors, but the measuring detector, which must be well calibrated obviously, yields the raw data as ***ground truth***. This is then the objective foundation, upon which discovery rests.

3. Community Debate and Setting Policy

We note that FAIR [5] is not the same as open data / science, it is however a prerequisite for it, nor does it impose, thus far, a criterion for data quality or trust. This point is where the International Union of Crystallography described the importance of quality and trust stating [6]:-

> *"that 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)."*

The IUCr in 2011 commenced a detailed examination of the practicalities of archiving the raw diffraction data sets in addition to the successively smaller processed and derived model datasets held in the crystallography databases [7]. Its 'Diffraction Data deposition Working Group' delivered its final report in 2017 at the World Congress in Hyderabad [8]. The conclusions were that this aspiration to preserve the raw diffraction data was both practical and also with numerous good reasons to do so within its fourteen recommendations.

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

4. Data policies and their implementations at photon and neutron facilities

The data policy of a research infrastructure describes the terms and conditions for how research data will be curated and treated. Essentially research infrastructure data policies describe three key aspects for scientific data:

(i) How a research infrastructure makes data Findable and Accessible FA(IR)?
(ii) How long data will be stored?
(iii) Who is responsible for curation of data?

For this article we surveyed the data policies of 34 photon and neutron Research Infrastructures. Our survey included most of the major synchrotron and neutron user facilities in North America, Europe and Asia Pacific regions. In each case we assessed the facility scientific data policy commitment for archival storage of raw data. The summary of this assessment is presented in Table 1.

Like in space science, primary data from photon and neutron sources is often automatically released to the public after an embargo period, during which access is privileged for the original researchers. This period is often 3 years reflecting the typical duration of a PhD programme. Many data policies mandate open access after the embargo period. Seventy percent of the facilities have a specific clause guaranteeing long term preservation of data. Of those facilities over half gave a specific time scale for data storage, ranging from 1 - 10 years. The significant consequence is there being no guarantee for longer term (> 10 y) curation of primary research data, which should be a concern for the scientific community.

Regarding this tensioning of resources, it is important to note the budget pressure for archival storage is heavily dependent upon data rates. There is a clear demarcation between photon and neutron facilities where the data rates of the latter are at least an order of magnitude less. Neutron facilities have historically been able to store all raw data collected. For the long term, achieving it is worth noting that due to the exponential increase in data rates and volumes over time the cost of storing data that has exited an embargo period is dwarfed by the cost of storing the

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

data for the current years.

Table 1 Result from the survey of data policies at Photon and neutron (PaN) facilities

Organisation	Policy defined data retention period	Embargo Period preceding open access	Ref
ORNL	Dependent upon data volume	-	https://tinyurl.com/y9wrb463
Argonne APS	No guarantee for archival storage of data	-	https://tinyurl.com/3btw54p5
BNL NSLSII	1 year	-	https://tinyurl.com/eunup24e
NIST NCNR	Not specifically defined	None or 18m	https://tinyurl.com/3spkpza8
SLAC	Responsibilities of facility users	-	https://tinyurl.com/2yzzz487
SPring8	No Online Data policy information	-	
Sirius	No Online Data policy information	-	
SSRF	No Online Data policy information	-	
JPARC MLF	Not specifically defined	3 years	https://tinyurl.com/vj5u5rsm
ANSTO Australian Synchrotron	yes 12m or 36m	Public after 36m	https://tinyurl.com/3az3bk75
Diamond Light Source	yes 30 days & long-term archive	3 years	https://tinyurl.com/nc2uwdu6

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

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
EUXFEL	5 years minimum (separate policy)	3 years	https://tinyurl.com/zp6yjebh , https://tinyurl.com/2cmb8cjc
PSI	5 years minimum, 10 years expected	3 years	https://tinyurl.com/rmc4naj
MaxIV	3 months	Not Specifically defined	https://tinyurl.com/2bm53zc6
ESS	No Online Data policy information	-	
Elettra	5-10 years	3 years	https://tinyurl.com/3vp73tvr

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

Alba	5 years	3 years	https://tinyurl.com/usb59c9m
Sesame	Minimum 5 years	3 years	https://tinyurl.com/sm8fwa3z
PaNData Policy Framework	10 years	3 years	https://tinyurl.com/28rwdyjd
PaNOSC Data Policy framework	10 years	3 years	https://tinyurl.com/tw9hju5a

Of the surveyed facilities with a specific policy for scientific data 15 facilities have specific terms for open access to data. Approximately a third of facilities provided online access to raw data or online access to the facility metadata catalogue. Two facilities, the NIST Centre for Neutron Research (NCNR), and the European Synchrotron Radiation Facility (ESRF) allow anonymous access to their data catalogue and to download raw data either immediately or after an embargo period. Other facilities require accessors to have an existing account i.e. they have to be known to the facility and accepted the terms of a data usage policy. The policies of European facilities reflect an ongoing commitment of the European Union towards open access to publicly funded research data. The UK (although having recently left the EU) has pioneered data archiving and openness after embargo period at ISIS and Diamond. ISIS, has achieved full data archiving and openness after an embargo period. Diamond has achieved full data archiving, the only synchrotron that has done so to our knowledge.

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. Thus data policies are often implemented in a staged approach.

The data policies show that research infrastructures appreciate the need for archival storage of data, along with the need for implementation of FAIR data management practices. It is often the case that realisation of this objective is blocked by an immediate tension of resources between the needs of provisioning instrumentation, beam days and critical supporting scientific infrastructure.

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

The majority of Photon and Neutron 'PaN' facilities in Europe have adopted a data policy based on the PaNdata data policy [9] and more recently on the PaNOSC data policy [10]. This is a big step for any institute because it involves a change of policy and needs the input of the user scientists and approval of the facility management and its governing bodies. However even if the tasks involved in setting up a metadata catalogue and data repository have been made much easier today with the availability of rich vocabularies, open source metadata catalogues, low cost high capacity storage systems and a rich collection of resources on how to implement FAIR data, many photon institutes struggle to implement a data policy for FAIR data. This is due to multiple factors, the main one being the lack of dedicated data curators in the institutes and the effort required to connect and adapt the databases to the local data acquisition and data curation workflow. An example of a community data catalogue is the Coherent X-ray Imaging and Diffraction Data Bank (CXIDB) [11].

Despite an often 3-year long embargo period for privileged use by the original PIs, researchers in many domains have a limited toolkit to assist in making their research data open and FAIR. This, together with a lack of established routes to properly acknowledge and reward original data producers in publications arising from reuse of data, creates a lethargy in the move towards an open research landscape. This results in a challenge for centralised facilities to generate metadata of sufficient quality for targeted cataloguing and re-use.

Another aspect of long-term preservation of raw experimental data is retention of data formats. The existence of NeXus as a data storage backend and vocabulary is an advantage for the photon and neutron communities. NeXus uses HDF5 as the low level data container, which is a widely adopted, efficient high performance format and corresponding library. HDF5 is backed by a non-profit company HDF Group. NeXus add the community governance of a domain specific structure and vocabulary, which allow a widely supportable complete description of photon and neutron measurements. Recently new additions like the Gold Standard for structural biology data have enhanced the NeXus definitions [12].

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

5. Case studies

Our case studies of data reuse in this article will illustrate our central premise that there are technically viable and financially affordable solutions to both, the data avalanche challenge and the challenge of simplifying the generation of FAIR data. Challenges that if left unmitigated threaten the efficiency and impact of the scientific workflow, be they at synchrotron, laser, neutron, as well as cryoEM facilities now also installed at synchrotron sites. This would also preclude credible introduction of future new technologies such as artificial intelligence and machine learning.

We assert that if data producers and scientists adopt a detailed strategic plan led by internationally agreed standards, centralized facilities, data managers, and publishers together, then large data volumes can be stored and shared efficiently and intelligently. This will preserve trust in the scientific process and enable novel re-use of data using emerging technologies

Long-term storage

The needs of the long term storage are not only increasing at a phenomenal pace for scientific data but for commercial applications too as tape is considered safer than disk for short and long term storage. The current roadmap for LTO tape storage predicts a promising 10-fold increase in tape capacity over the next 10 years (see figure 2) which combined with new data compression algorithms being developed, should enable facilities to archive all raw data in the future.

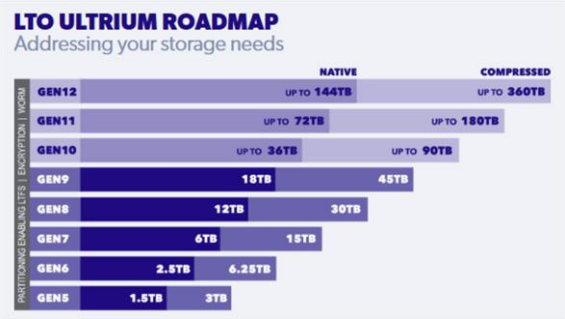


Figure 2 Roadmap for tape storage capacity. GEN8 is the current standard in 2021 (source: https://www.lto.org/roadmap/).

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

Case Study 1 - The macromolecular crystallography 'MX' current state of the art: the examples of COVID 19 and the anti cancer platins data reuse".

In the medical pandemic of the coronavirus COVID-19, there has been intense research activity for seeking both a vaccine and an inhibitor drug around the world. In the lockdowns by governments many laboratories, including X-ray, electron and neutron facilities, stayed open [13] or at least maintained a core minimum operation. The 3D atomic resolution cryocrystal and cryoEM structures and neutron room temperature structures gradually have been deposited in the Protein Data Bank (PDB). Many included the associated primary experimental data uploaded at the general archives such as Zenodo. From the PDB depositions a website of data resources has been established whereby a re-refinement of the crystal structures including their ligands was evaluated and where necessary the authors PDB deposits improved upon [14,15] . This approach thus far has largely followed the paradigm where the raw diffraction data were not available. Access to the raw diffraction data processing would allow to verify the choice of crystal space group, check whether the diffraction resolution was chosen correctly, and assess if there was diffuse scattering evidence of dynamics of the protein structure i.e. its plasticity and flexibility. In a different medical theme, the anti-cancer platins data and data reuse, more than 30 raw diffraction data sets were shared from the gold open access research articles, each with PDB depositions as well. This formed a comprehensive raw data sharing and led to improvement in one case in the diffraction data resolution by reusers of the data and then again by JRH's laboratory [16]. These two examples in this case study document the MX situation for raw data reuse as a mixture of reuse of raw data and of corresponding PDB data deposition. The PDB has also very usefully introduced space for the raw data doi in a deposit and for versioning of a PDB deposit by the depositors as it may improve.

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

Case Study 2 – Paleontology data from 3rd generation photon sources available as Open Data producing new publications.

Palaeontology has seen a big boost from 3rd generation photon sources and many publications and datasets have been generated over the last 20-30 years. The advantages of non-destructive high resolution tomography have been a major application for many paleontological samples e.g. the Sediba skull, which was judged one of the major scientific advances of the last decade by the Smithsonian Institute. Due to the large number of samples in some collections, datasets exceed the number of scientists available to analyse them. To address this, the ESRF setup a database [17] with many datasets which are now publicly available. Since the creation of the database roughly 30 publications not involving the original authors have been made according to the main scientists in charge of the database.

6. Conclusions and summary

In summary, leading photon and neutron research facilities in the world offer cutting edge experiments but are also at the forefront of scientific data management. Behind this are two driving forces. Firstly, the scientific community itself, whose trustworthiness has been called into question through a perceivable lack of reproducibility. The scientists' shield to such wounding criticism is through the attachment of their data (raw, processed, and derived) to their publications, a methodology championed by astronomers and crystallographers as well as summed up by the FAIR (Findable, Accessible, Interoperable and Reusable) movement. Secondly, the funding agencies, in their response to governments and taxpayers, seek faster discoveries and if possible better value for money. Thus, data should be released for use beyond the original research team.

The physical data archiving need of the scientists, is much less daunting to the facilities' data management. It only mandates preservation of the exploited subset of the collected data volume. The third-party use, favoured by governments and their funding agencies, does assume storage of all raw data.

The facilities' users gain a major advantage of the professional management by the facilities or publishers. The bottom line is that the colossal expansion of the data archives presents great opportunities to all scientists including users of the photon and neutron facilities. This data archive expansion of capability is continuing at pace.

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

The objectivity that science promises is within reach when the as-measured-experimental-data can be preserved and linked to the subjective narratives of authors.

Acknowledgments

The PaNOSC project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No 823852. JRH is very grateful to the University of Manchester for its provision to its researchers of research data archiving and dataset registration, not least during a transition these last five years or more of the facilities to raw data preservation and digital object identifier registration.

References

1. Baker, M (2016) 1,500 *Scientists lift the lid on reproducibility, Survey sheds light on the 'crisis' rocking research* **in** Nature 533, 452–454.
2. Physics Today (September 2020) *Synchrotrons face a data deluge* https://physicstoday.scitation.org/do/10.1063/pt.6.2.20200925a/full/
3. Allea (All European Academies) (2017) *The European Code of Conduct for Research Integrity Revised Edition* http://www.allea.org/wp-content/uploads/2017/03/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017-1.pdf
4. Directorate-general for research and innovation (European Commission) and pwc eu services (2016) *Cost-benefit analysis for fair research data, cost of not having fair research data* https://doi.org/10.2777/02999
5. M. D. Wilkinson et al (2016) Comment: *The FAIR Guiding Principles for scientific data management and stewardship* Scientific Data | 3:160018 | https://doi.org/10.1038/sdata.2016.18
6. Hackert, M L, van Meervelt, L , Helliwell, J R, and McMahon, B *Open data in a big data world: a position paper for crystallography* https://www.iucr.org/iucr/open-data
7. Ian Bruno, Saulius Gražulis, John R Helliwell, Soorya N Kabekkodu, Brian McMahon and John Westbrook (2017) *Crystallography and Databases*. Data Science Journal. 16, p.38, 1-17 https://doi.org/10.5334/dsj-2017-038
8. IUCr Diffraction Data Deposition Working Group (DDDWG) Final Report https://www.iucr.org/resources/data/dddwg/final-report by John Helliwell (UK), Steve Androulakis (Australia), Sol Gruner (USA), Loes Kroon-Batenburg (Netherlands), Brian McMahon (UK), D. Marian Szebenyi (USA), Tom Terwilliger (USA), Edgar Weckert (Germany), John Westbrook (USA) and Heinz-Josef Weyer (sadly deceased, Switzerland)
9. Dimper, Rudolf (2011) *Common policy framework on scientific data* https://doi.org/10.5281/zenodo.3738497
10. Andy Götz, Jean-Francois Perrin, Hans Fangohr, Daniel Salvat, Florian Gliksohn, Anders Markvardsen, Abigail McBirnie, Alejandra Gonzalez-Beltran, Jonathan Taylor, Brian Matthews, (2020) *PaNOSC FAIR research data policy framework*, https://doi.org/10.5281/zenodo.3826039
11. Maia, F. R. N. C. *The Coherent X-ray Imaging Data Bank* Nat. Methods S 9, 854–855 (2012).
12. Herbert J. Bernstein, Andreas Förster, Asmit Bhowmick, Aaron S. Brewster, Sandor Brockhauser, Luca Gelisio, David R. Hall, Filip Leonarski, Valerio

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

Mariani,g Gianluca Santoni,Clemens Vonrhein and Graeme Winter *Gold Standard for macromolecular crystallography diffraction data* IUCrJ 7, 784-792 and references therein.

13. Kramer, D. (2020) *World's physics instruments turn their focus to COVID-19* Physics Today https://physicstoday.scitation.org/doi/10.1063/PT.3.4470
14. Wlodawer, A, Dauter, Z, Shabalin, I G, Gilski, M, Brzezinski, D, Kowiel, M, Minor, W, Rupp, B. Jaskolski, M (2020) *Ligand centred assessment of SARS-CoV-2 drug target models in the Protein Data Bank*. The FEBS Journal 287 (2020) 3703–3718.
15. M. Jaskolski, Z. Dauter, I. G. Shabalin, M. Gilski, D. Brzezinski, M. Kowiel, B. Rupp and A. Wlodawer *Crystallographic Models of Sars-Cov-2 3clpro: in-depth assessment of Structure Quality and Validation* IUCRJ (2021) 8, 238-256.
16. Simon W. M. Tanley, Antoine M. M. Schreurs, Loes M. J. Kroon-Batenburg and John R. Helliwell *Re-refinement of 4g4a: room-temperature X-ray diffraction study of cisplatin and its binding to His15 of HEWL after 14 months chemical exposure in the presence of DMSO* Acta Cryst. (2016). F72, 253–254.
17. ESRF heritage database for palaeontology, evolutionary biology and archaeology (http://paleo.esrf.eu)

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