Linking X-Ray Diffraction Data to Our Publications Allows Objectivity in Our Science

John R Helliwell

Chairman of the Committee on Data of the

International Union of Crystallography (IUCr)

And IUCr Representative to CODATA

November 22nd 2022

Talk Contents

- Trust in science
- Opportunities linking raw diffraction data to our publications
- Consulting the IUCr Commissions
- Unpublished data
- Modern data rates
- Some topical questions on raw diffraction data preservation/release
- Can traditional peer review of article with data used by IUCr Journals be applied to databases or facility data archives?
  - Collaboration with PDBj and its XRDa raw diffraction data archive (ongoing)
  - Synchrotron, XFEL and Neutron facilities' data catalogues
- Conclusions

Trust in Science

- "**FAIR**" is a very widely used acronym in science meaning data need to be ***Findable, Accessible, Interoperable and Reusable*** [M. D. Wilkinson et al (2016) *Comment: The FAIR Guiding Principles for scientific data management and stewardship Scientific Data* | 3:160018 | DOI: 10.1038/sdata.2016.18]

- National Academies of Sciences, Engineering, and Medicine 2019. ***Reproducibility and Replicability in Science*** Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

For a scientist, trust is not blind or uncritical, and the availability of underpinning data is essential in revisiting a study (and so forming a judgement on the level of reliability).

The Crystallographic Information Framework[#] facilitates trust in crystal structures

Trust is needed in:

- Data transmission/exchange
  - Crystallographic Information File (1991)
- Data consistency
  - checkCIF for derived (coordinate) data (1998)
  - checkCIF including structure factors (2007)

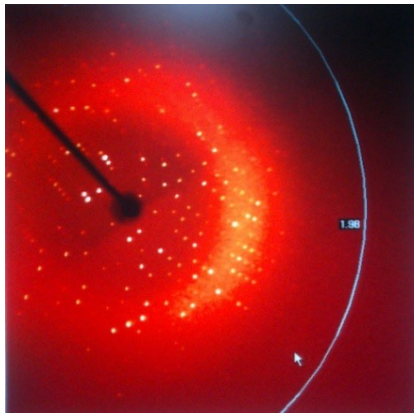
 # IUCr COMMITTEE FOR THE MAINTENANCE OF THE CIF STANDARD (COMCIFS)

In 2003 wwPDB Validation started, which IUCr keenly supported; "Validation Report" as a term was 2010 onwards

Our modern data zoo

Data can mean any or all of:

1. raw measurements from an experiment
2. processed numerical observations
3. derived structural information



A raw diffraction image; Thousands or more of these make a complete experimental raw diffraction dataset.

```
loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0      772.37      856.47     28.20 o
   4   0   0     1445.15     1446.80     39.55 o
   6   0   0     1130.79     1097.08     30.62 o
   8   0   0     1347.13     1490.27     55.41 o
  10   0   0     3273.01     3545.64    154.91 o
  12   0   0       48.20       40.50      4.56 o
  14   0   0       79.87       63.02      7.91 o
   2   1   0     2093.70     1975.83     47.36 o
   3   1   0    33795.10    34884.29   1287.71 o
   4   1   0     2298.16     2035.72     38.24 o
   5   1   0        9.73       36.06      5.59 o
   6   1   0      449.80      506.89     11.92 o
   7   1   0        1.81        7.91      5.59 o
   8   1   0       43.36       28.81      6.79 o
   9   1   0       64.18       48.51      6.02 o
  10   1   0     1412.22     1628.54     45.96 o
  11   1   0      242.68      279.96      9.70 o
  12   1   0       14.96       10.52      3.84 o
  13   1   0       16.87       15.76      4.56 o
  14   1   0       16.46        7.91      7.91 o
  15   1   0        0.00        3.95      5.59 o
   0   2   0     2443.71     2679.14     61.27 o
   1   2   0    23397.80    23770.90    546.30 o
   2   2   0    20572.37    19502.51    520.01 o
   3   2   0     8854.88     8282.53    169.57 o
```

Table 1. *Selected geometric parameters (Å, °)*

Fe1—C9	2.030 (4)	Fe1—C7	2.049 (4)
Fe1—C5	2.036 (3)	Fe1—C11	2.053 (4)
Fe1—C12	2.038 (4)	S1—C1	1.693 (3)
Fe1—C13	2.038 (4)	N1—C1	1.315 (4)
Fe1—C4	2.038 (3)	N2—C1	1.345 (4)
Fe1—C8	2.041 (4)	N2—N3	1.387 (4)
Fe1—C10	2.042 (4)	N3—C2	1.290 (4)
Fe1—C6	2.048 (4)		
C1—N2—N3	118.2 (3)	N2—C1—S1	120.1 (2)
C2—N3—N2	116.9 (3)	N3—C2—C4	115.6 (3)
N1—C1—N2	117.0 (3)	N3—C2—C3	125.2 (3)
N1—C1—S1	122.8 (2)	C4—C2—C3	119.3 (3)

Each includes metadata i.e. also are data.

Today we can start to include our raw data as part of our preserved workflow

Subjective

Science Article "Narrative"

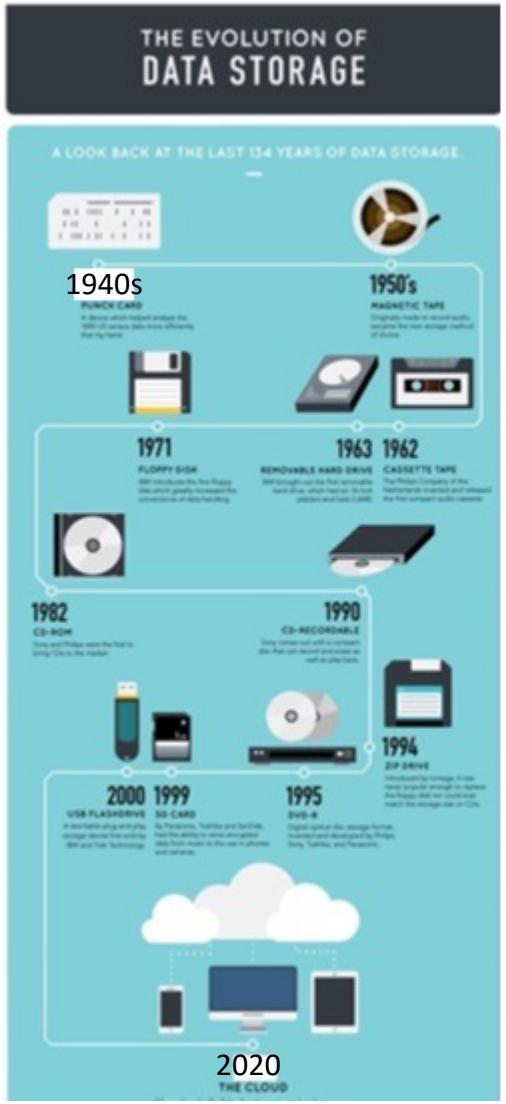
Derived Data

Processed Data

Raw Experimental Data (Ground Truth)

Objective[£]

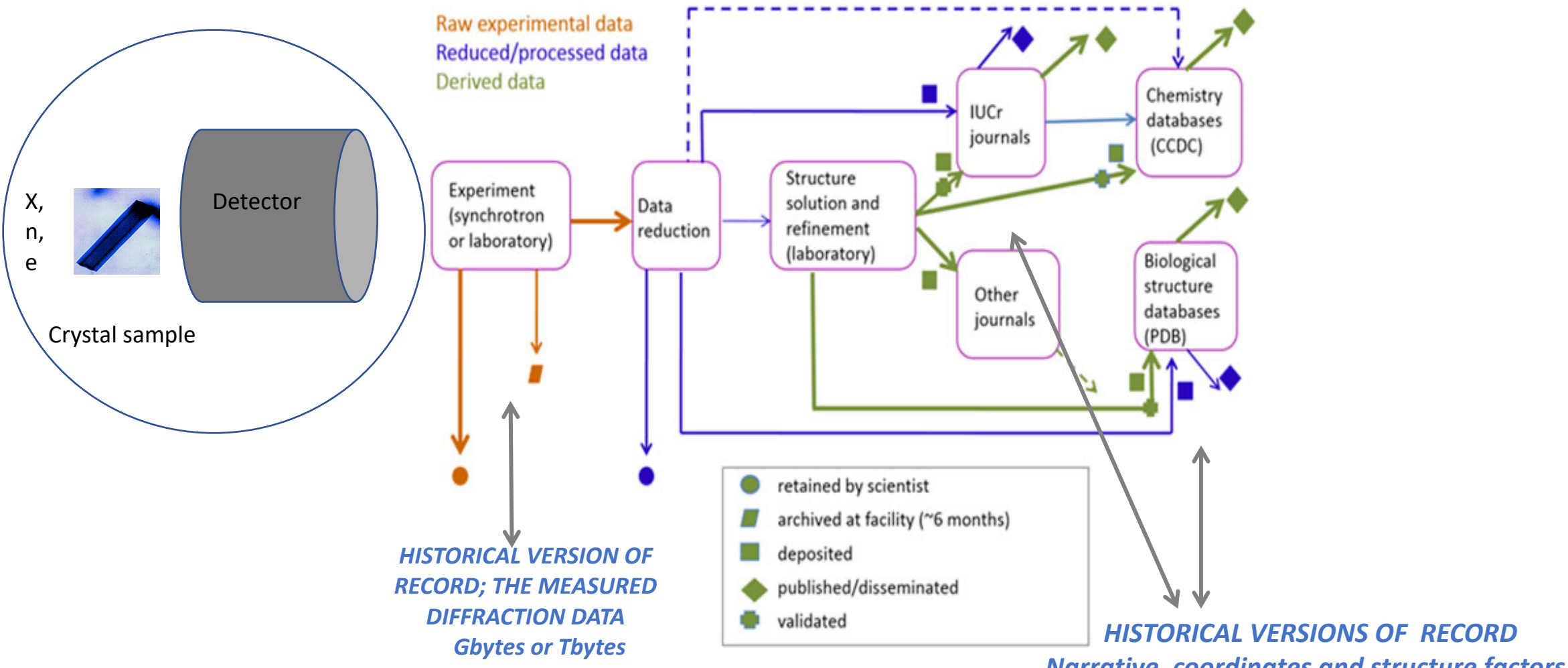
Digital data storage capability



[£] *Nb an instrument must be calibrated by a person and this leaves some degree of subjectivity*

Coherent approach of crystallography:

Crystallographic Information Framework (CIF) ontologies at each stage



A recent development (mid-2022)

https://journals.iucr.org/services/datasharingpolicy.html

IUCr Journals
CRYSTALLOGRAPHY JOURNALS ONLINE

home submit subscribe open access

data-sharing policy

The IUCr supports the principles of transparency and openness in scientific research. This means making research data available to facilitate the reproduction, validation, reuse and reinterpretation of research findings, leading ultimately to more effective research discovery, i.e. Findable Accessible Interoperable Re-usable (FAIR).

IUCr journals mandate data sharing and peer review of crystal structure data

The IUCr has adopted a data-sharing policy that requires the crystal structure data supporting the results in an article to be peer reviewed and archived either with the IUCr or in an appropriate public repository. Full details of the requirements are given in the Notes for authors. For articles describing scripts or program code, whenever possible these and any other documentation necessary for reproduction of the published results should also be publicly archived. This policy applies to all IUCr journals. (In rare cases exceptions may be granted by the editors, for example when the sharing of data may compromise ethical standards or legal requirements.)

Published articles contain links to the data either on the IUCr site or in the repository together with links to other supporting information. All relevant accession, reference or identification codes, or other persistent identifiers such as DOIs are included in articles.

It is the practice of IUCr journals to provide free access to all supplementary materials and supporting data files deposited with a published article. When the data are reused, we ask that proper attribution is given to the associated source article.

DATA REPOSITORIES

The IUCr recommends that authors deposit data sets in standard, subject-specific public repositories. A list of repositories relevant to crystallographic data is available at https://www.iucr.org/resources/data/databases. Additional repositories may be found by visiting https://www.re3data.org or https://fairsharing.org, which list registered and certified data repositories.

DATA REFERENCES

Identification of individual structures in an article by use of database codes should be accompanied by a full citation of the original literature in the reference list.

Where authors have used a publicly available data set from a repository this should be cited in the published article according to the following style:

Authors (Year). Data set title. Persistent identifier.

An example citation is:

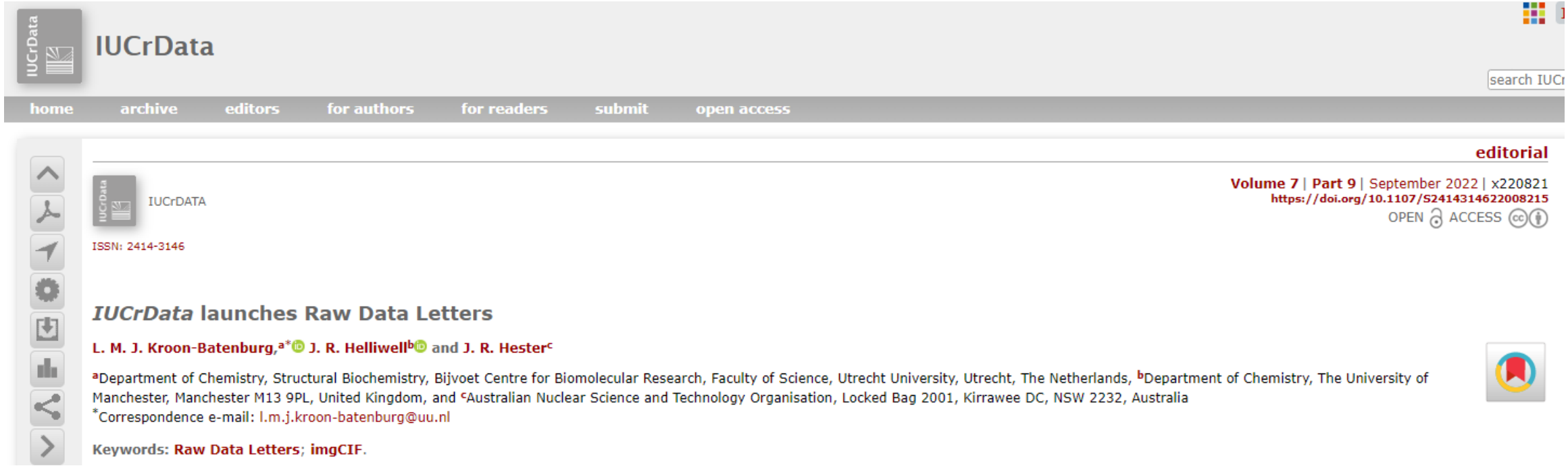
e.g. → Brink, A. & Helliwell, J. R. (2019). Raw diffraction images. Formation of a highly dense tetra rhenium cluster in a protein crystal and its implications in medical imaging. https://doi.org/10.5281/zenodo.2874342

New opportunities and initiatives stemming from being able to store large quantities of raw data

- *Better understanding of what we do experimentally*
- *Harnessing new methods and software*
- *Enabling new science*
- *Understand the subjective choices made in data processing*

- *The above are in addition to the decades long benefits of:-*
  - *archived coordinates (structure and bonding trends and snapshots of conformational dynamics)*
  - *then processed diffraction data (re-use/re-refinement of a structure based on authors' SFs)*

IUCr Journals has launched IUCrData's Raw Data Letters>>>



IUCrData

search IUCr

home archive editors for authors for readers submit open access

editorial

IUCrDATA

Volume 7 | Part 9 | September 2022 | x220821
https://doi.org/10.1107/S2414314622008215

OPEN ACCESS

ISSN: 2414-3146

IUCrData launches Raw Data Letters

L. M. J. Kroon-Batenburg,[a]* J. R. Helliwell[b] and J. R. Hester[c]

[a]Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands, [b]Department of Chemistry, The University of Manchester, Manchester M13 9PL, United Kingdom, and [c]Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia
*Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Keywords: Raw Data Letters; imgCIF.

IUCrData

ISSN 2414-3146

Received 20 April 2021
Accepted 1 May 2021

Keywords: twinning; diffuse scattering; tetraspanin $CD9_{EC2}$.

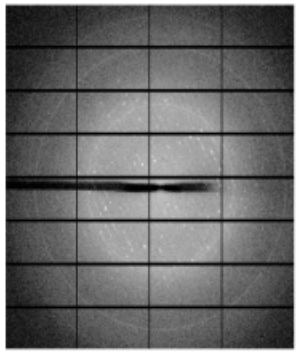
Second extracellular domain of human tetraspanin CD9: twinning and diffuse scattering

Viviana Neviani, Martin Lutz, Wout Oosterheert, Piet Gros and Loes Kroon-Batenburg*

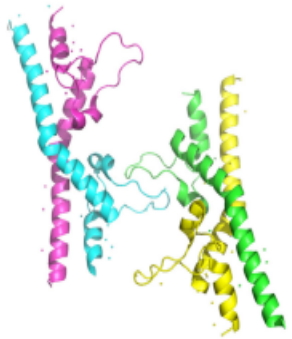
Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands. *Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Remarkable features are reported in the diffraction pattern produced by a crystal of tetraspanin $CDCD9_{EC2}$, the structure of which was described previously [Oosterheert *et al.* (2020). *Life Sci. Alliance*, **3**, e202000883]. $CD9_{EC2}$ crystallized in space group *P*1 and was twinned. Concurrent with the twinning, diffuse streaks were seen in the direction perpendicular to the twinning interface. Preliminary conclusions are made on packing disorder and potential implications for the observed molecular structure. It is envisaged that the raw diffraction images could be very useful for methods developers in trying to remove the diffuse scattering to extract accurate Bragg intensities or by using it to model the effect of packing disorder on the molecular structure.

Raw data



Structure



Raw diffraction data

HDF5 data file, DOI: https://doi.org/10.5281/zenodo.1234567

Metadata ImgCIF file, DOI: https://doi.org//10.1107/S2414314622000384/me6134.cif

checkImgCIF report
[CheckCif for Raw Data]

```
ImgCIF checker version 2022-07-16

Checking block 5886687 in he4557img.cif

Running checks (no image download)
======================================

Testing: Required items: PASS

Testing: Data source: PASS

Testing: Axes defined: PASS

Testing: Our limitations: PASS

Testing: Detector translation: PASS

Testing: Scan range: PASS

Testing: All frames present: PASS
   All frames present and correct for SCAN1

Testing: Detector surface axes used properly: PASS

Testing: Pixel size and origin described correctly: PASS

Testing: Check calculated beam centre: PASS

Testing: Check principal axis is aligned with X: PASS
Testing presence of archive:

Testing: All archives are accessible: PASS

Running checks with downloaded images
======================================


Testing image 4: Image type and dimensions: PASS

Testing image 4: Overloaded values present: PASS

====End of Checks====
```

Raw data table generated from the CIF

Raw data	
DOI	https://doi.org/10.5281/zenodo.5886687
Data archive	Zenodo
Data format	HDF5
Data collection	
Beamline	Diamond I04
Detector	
Temperature (K)	
Radiation type	Synchrotron X-ray source
Wavelength (Å)	0.979491
Beam centre (mm)	−166.874, 172.497
Detector axis	−Z
Detector distance (mm)	-287.22
Swing angle (°)	
Pixel size (mm)	0.075 × 0.075
No. of pixels	4148 × 4362
No. of scans	1
Exposure time per frame (s)	
Scan axis	ω, X
Start angle, increment per frame (°)	0.0, 0.1
Scan range (°)	360.0
No. of frames	3600

IUCrData

Accurate intensity integration in the twinned c-form of o-nitroaniline

Martin Lutz and Loes Kroon-Batenburg

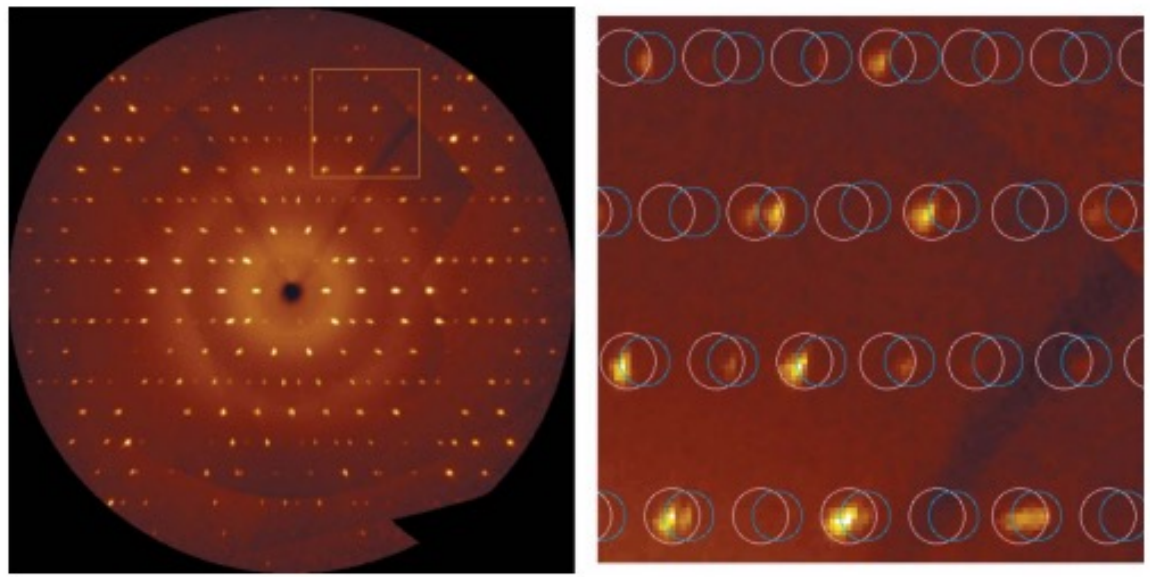


Figure 4
Left: simulated precession photograph in the *h*0*l* plane of (I) up to a resolution of 0.9 Å. The reconstruction is based on seven scans with a total of 3324 raw images. Right: zoomed image, is from the yellow square in the left image. White circles are the predicted impacts for the first twin component, blue circles for the second.

Are all areas of crystallography & diffraction the same in their raw data archiving needs?

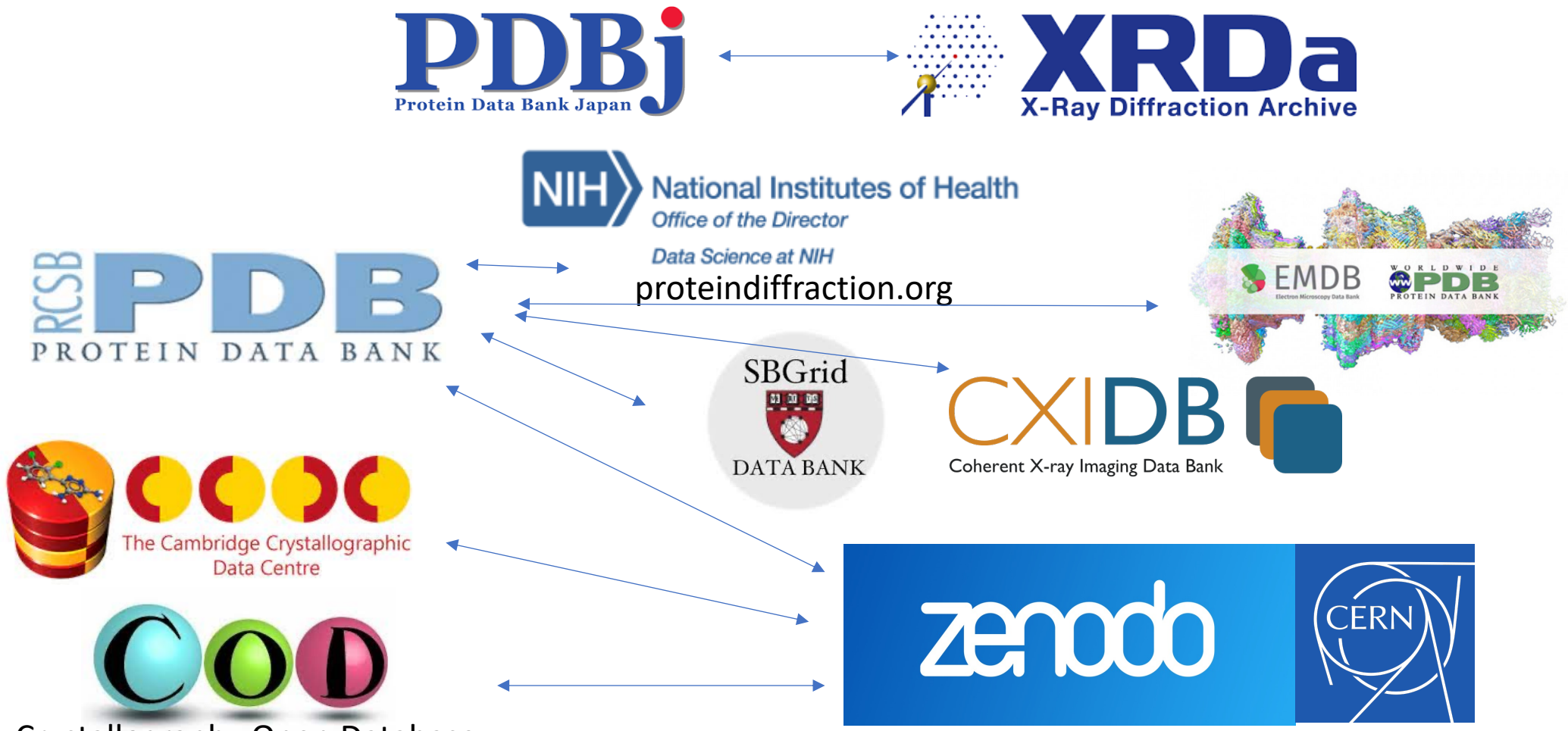
- IUCr Commission on Biological Macromolecules has effected changes in IUCr Journals Notes for Authors that ***data processing methods and new structures papers must have their underpinning raw diffraction data doi cited.***

- Chemical crystallographers organised a Workshop linked to IUCr Prague https://www.iucr.org/resources/data/commdat/prague-workshop-cx to examine the question ***When should small molecule crystallographers publish their raw diffraction data? Answer: in special cases***

- X-ray powder diffraction has a “policy discussion paper” in J Appl Cryst in 2018 by Miguel Aranda (ALBA Science Director until recently) ***Sharing powder diffraction raw data: challenges and benefits***

Open Science: publications and data

- Link all the underpinning data to the publication, raw, processed and derived. Exemplars are for eg:



Crystallography Open Database

Examples from Europe's facilities

- Diamond Light Source, ESRF and Soleil save all measured data and have a policy committed to release of all raw data after 3 years

- Pioneering from 2018, for 2 years ESRF have generated one DOI per proposal using DataCite (examples: https://search.datacite.org/works?query=10.15151%2F*), users can also create additional DOI per dataset using the ESRF data portal. ESRF asks their users to provide their DOI of the data in their scientific articles.

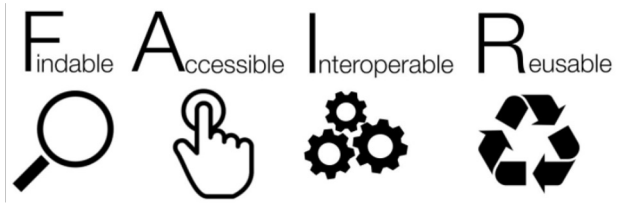
- In Germany there is the National Forschungsdateninfrastruktur (https://www.nfdi.de/) bringing proper data management tools and metadata harvesting to many science areas including the photon and neutron sciences (***DAPHNE4NFDI, DAten aus PHotonen und Neutronen Experimenten***).

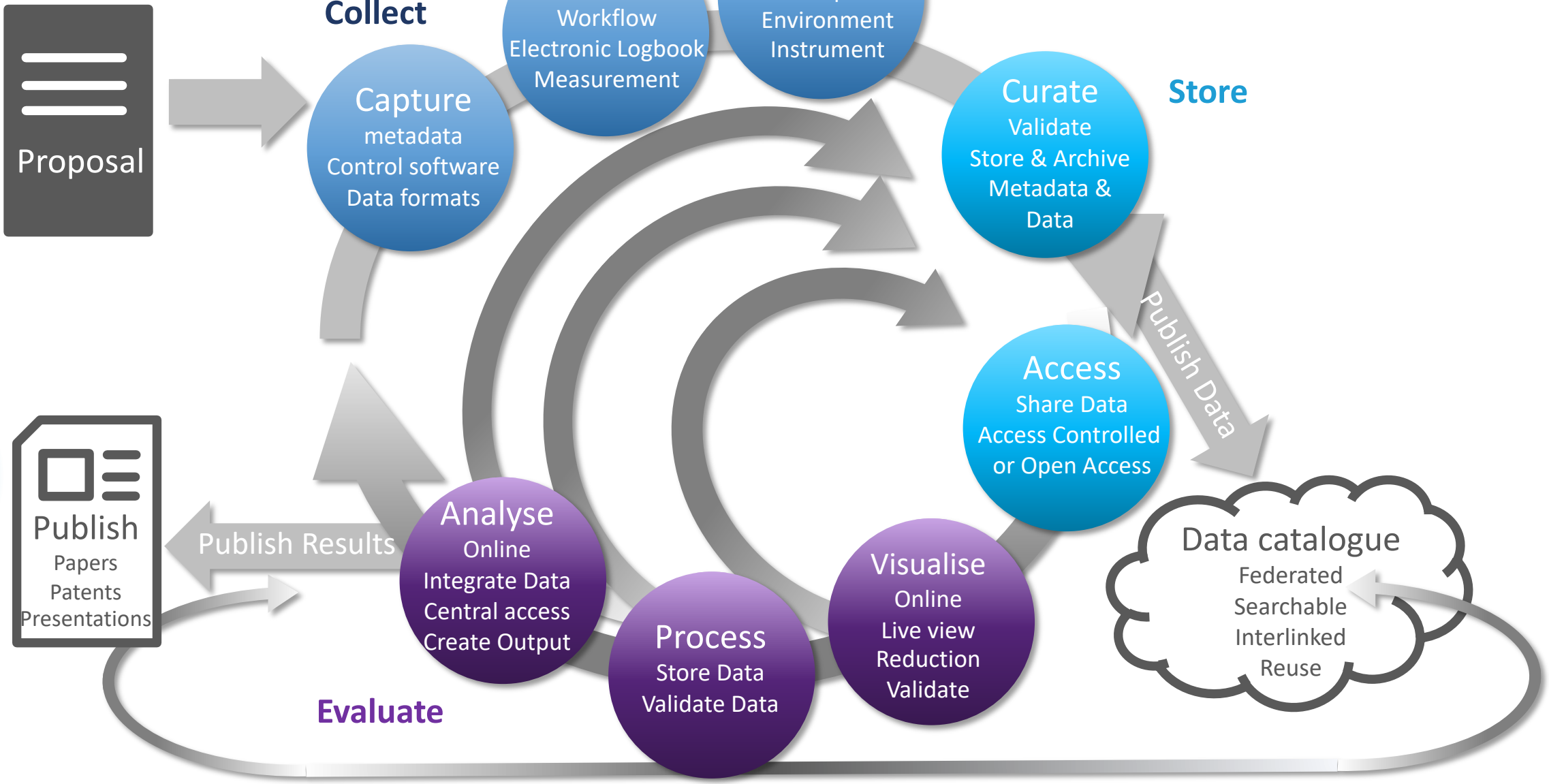
- A coordinated European Open Science Cloud is imminent, to which PANOSC is affiliated (Photon and Neutron Open Science Cloud)

DAPHNE Research Data Management Strategy

www.DAPHNE4NFDI.de

Slide kindly from Bridget Murphy,U Kiel





Open Science as a Grand Principle?

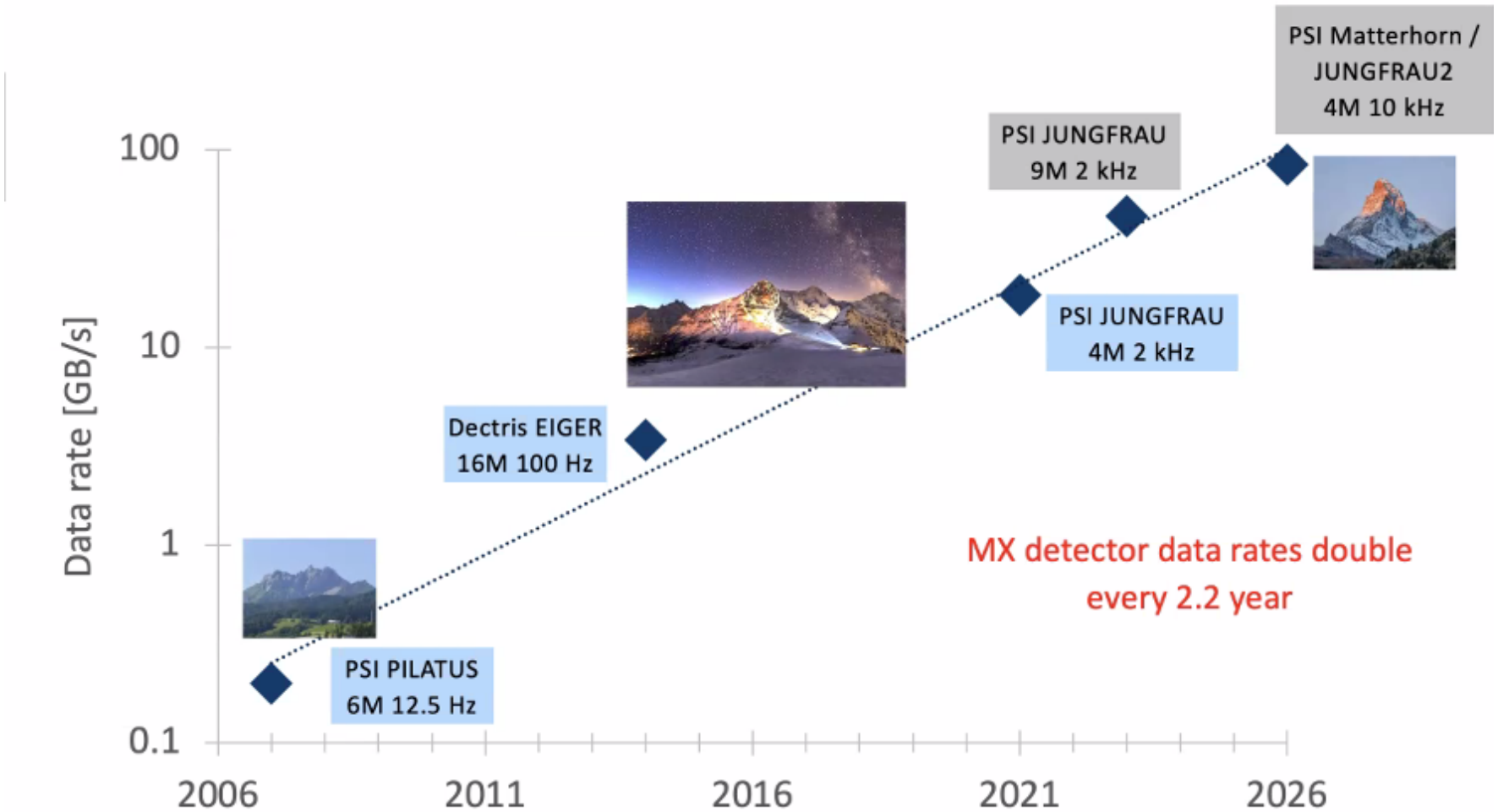
- Grand Principle: ***All measured data, including even unpublished raw data with no derived molecular structure model, should be made open***
- ***Certainly useful would be such as:***
  - ***ESRF Paleontology initiative; Researchers agree that they cannot analyse all the data unless they make it open***
  - ***Diamond Light Source covid research initiative; raw data are copied to Zenodo by the researcher***

- ***Unuseful experimental data are empty data frames such as when the beam fails to hit a crystal; so, delete.***

- ***Grey area for me: automatic release after 3 years even when no publication***

PAUL SCHERRER INSTITUT PSI

Future challenges

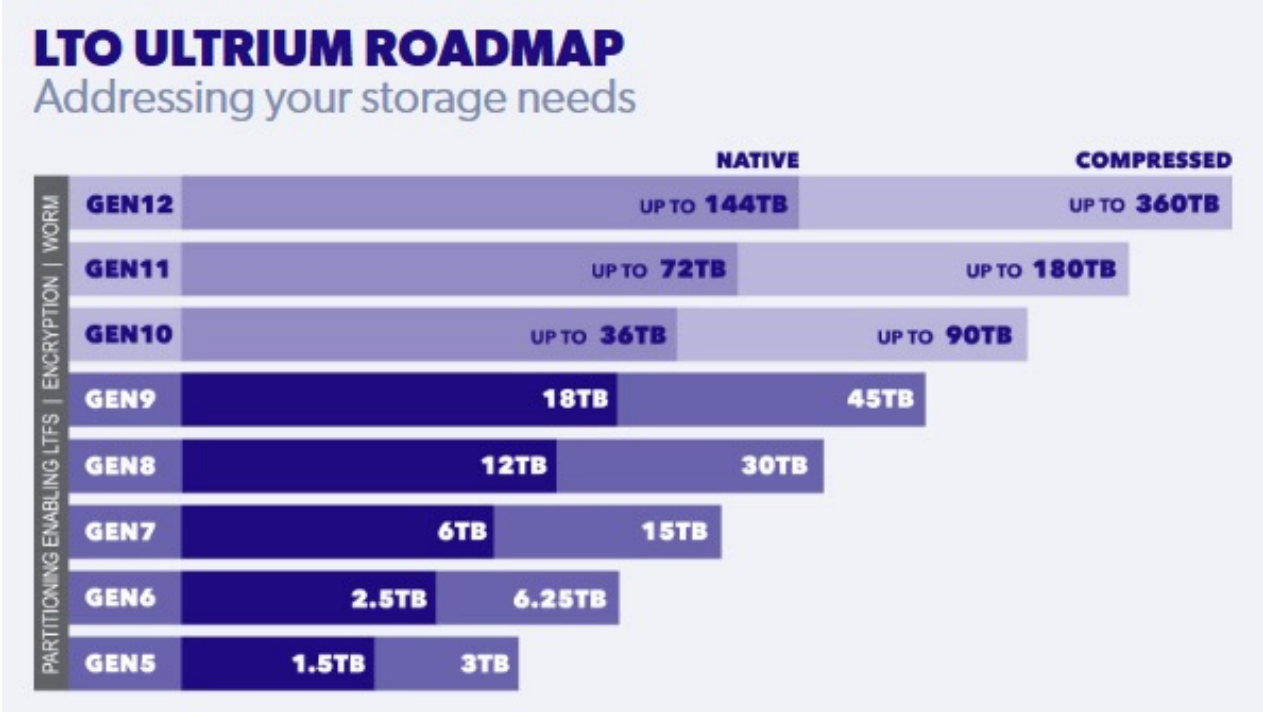


Presented here with the permission of Filip Leonarski, PSI

[from the *High Data Rates in Macromolecular Crystallography Workshop* April 2022, Organised by Herbert Bernstein]

What about tape storage capacities?

GEN8



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

From: *The vital role of primary experimental data for ensuring trust in (Photon & Neutron) science*
Paper written for the PaNOSC project by Götz , Helliwell , Richter, Taylor (2021) 10.5281/zenodo/5155882

With such data rates, facilities are providing at-facility-raw-data-processing. This also provides clarity about the workflow that was used for a particular project. If a doi is provided then a publication can reference those raw and processed data files.

Examples:-

https://www.panosc.eu/news/new-video-released-on-visa-virtual-infrastructure-for-scientific-analysis/

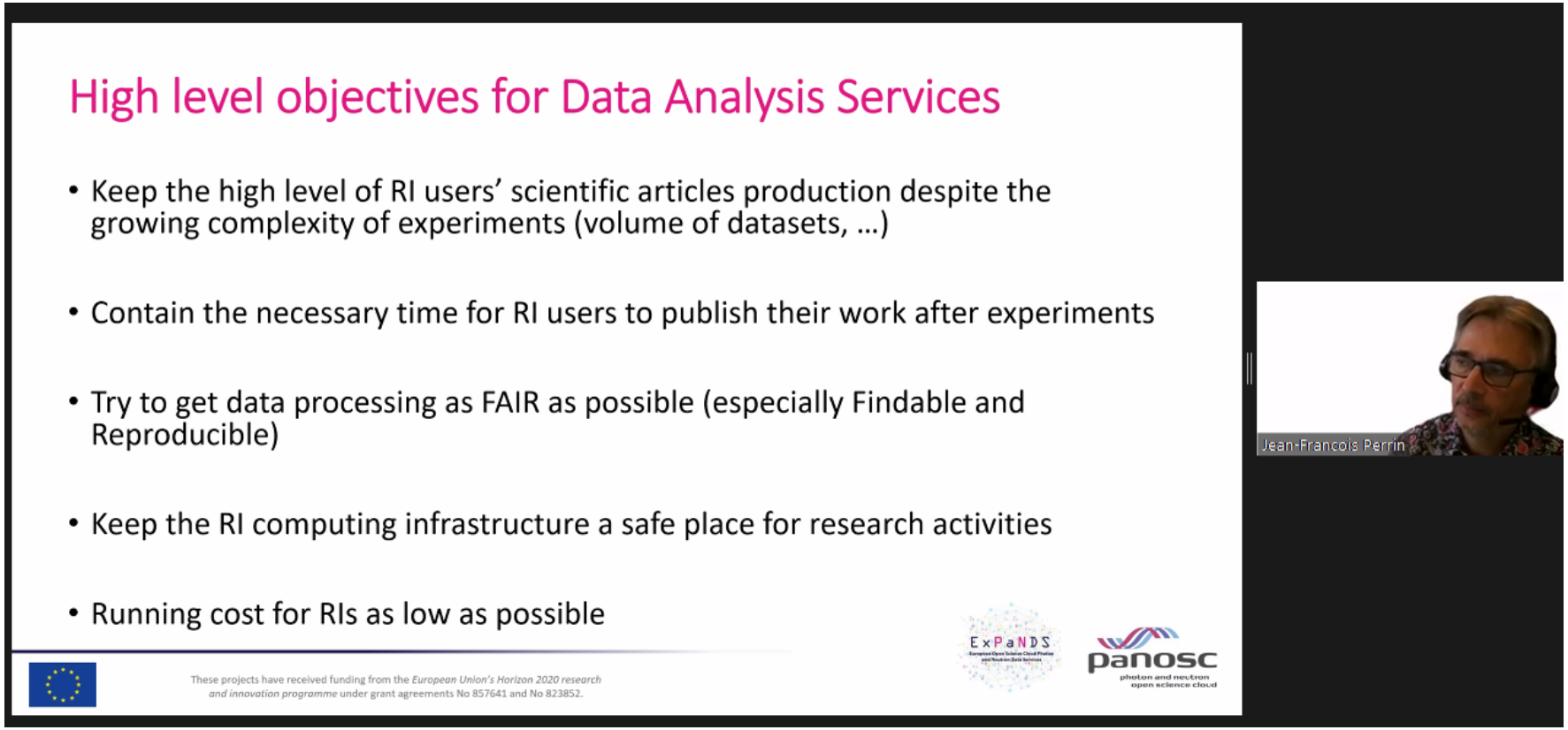
CCP4: Software for Macromolecular X-Ray Crystallography

ExPaNDS=European Open Science Cloud Photon and Neutron Data Services

ExPaNDS

panosc
photon and neutron open science cloud

As explained at the recent PANOSC training day for VISA by Jean-Francois Perrin of ESRF EBS:-



Presented here with the permission of Jean-Francois Perrin, ESRF EBS.

Actually, we need to understand better:-

Q1. What fraction of measured raw data leads to publication? *i.e.* if every measured data frame leads to publication, we have a ‘permanent’ storage, and costs, challenge.

Q2. If a publication doesn’t work out **who should decide those data can be deleted? and after how many years? [Presumably the entity bearing the costs decides.]**

Q3. Maybe, in due course different areas of science could/would **reach maturity** in the same way as chemical crystallography, *i.e.* where raw data are not vital to be preserved in every experiment?

How to approach Question 2?

- The IUCrData Raw Data Letters, amongst various possibilities, can allow a PI to explain why an analysis is taking longer than (say) 3 years. The PI might be under such a (legitimate) pressure from their Facility.
- Alternatively it could be agreed between the PI and their Facility to delete a data set, even where a DOI had been previously assigned. If so then:-
  - *Might it be fruitful to start thinking about standard ways of annotating a DOI on deletion of the data set to declare why it was deleted?*

Open Discussion:
There is the IUCr CommDat Forum for Public Inputs on Data

FAQ | Search

Board index / Standing Committees and Working Groups / Public input to CommDat

Public input to CommDat

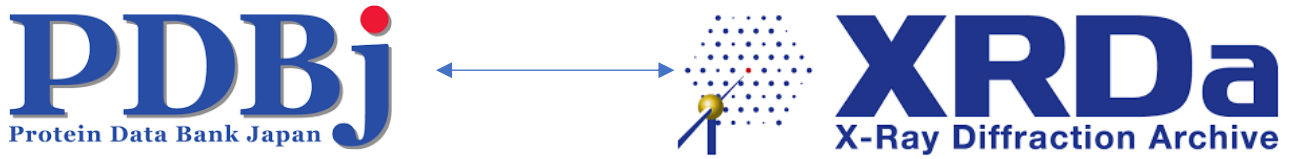
New Topic | Search this forum... | 43 topics | 1 | 2

TOPICS	STATISTICS	LAST POST
The 4Rs and crystal structure analysis: teaching and education article in J Appl Cryst by JRH » Tue Aug 30, 2022 8:34 am	Replies: **0** Views: **1064**	by JRH Tue Aug 30, 2022 8:34 am
start of Raw Data Letters section in IUCRData by Loes Kroon-Batenburg » Wed Aug 24, 2022 1:59 pm	Replies: **0** Views: **455**	by Loes Kroon-Batenburg Wed Aug 24, 2022 1:59 pm
USA NASEM, USNCCr and NIST "Crystallography and Structural Databases" Course by JRH » Wed Apr 06, 2022 10:01 am	Replies: **0** Views: **4432**	by JRH Wed Apr 06, 2022 10:01 am
The (pre)history of CommDat by Brian McMahon » Thu Mar 31, 2022 5:30 pm	Replies: **0** Views: **2904**	by Brian McMahon Thu Mar 31, 2022 5:30 pm
2021 Annual Report to Executive Committee by Brian McMahon » Tue Mar 22, 2022 12:57 pm	Replies: **0** Views: **2975**	by Brian McMahon Tue Mar 22, 2022 12:57 pm
Just who does own research data? by JRH » Fri Feb 18, 2022 3:06 pm	Replies: **0** Views: **6585**	by JRH Fri Feb 18, 2022 3:06 pm

Can traditional peer review of article with data used by IUCr Journals be applied to databases?

The ***Diffraction Data Case Study for CODATA's GOSC*** of PDBj+XRDa

We focus on medically important proteins



John R Helliwell (UK) and Genji Kurisu (Japan)

with

Loes Kroon-Batenburg (The Netherlands)

Our activities

Deliverables

Reproducibility of data sets is paramount.

We aim for a single point of contact for definitive molecular models, namely at the PDBj and its XRDa, the X-ray Diffraction Data Archive based at the Institute of Research in Japan.

We aim to avoid dispersed multiple versions of a protein model derived from a single raw diffraction data set. Controlled versioning procedure of PDB entries should be tightly linked.

A critical deliverable is to realise metrics of ‘definitive reusability’ which would then be applicable to the individual diffraction data sets held in the XRDa. These metrics and the definitive diffraction data files are a bedrock of interoperability.

Achievements and progress

We have provided assessments of ***medically relevant protein crystal structure*** deposits into PDBj which have XRDa raw diffraction data equivalent data files.

We have presented our work thus far at the:-

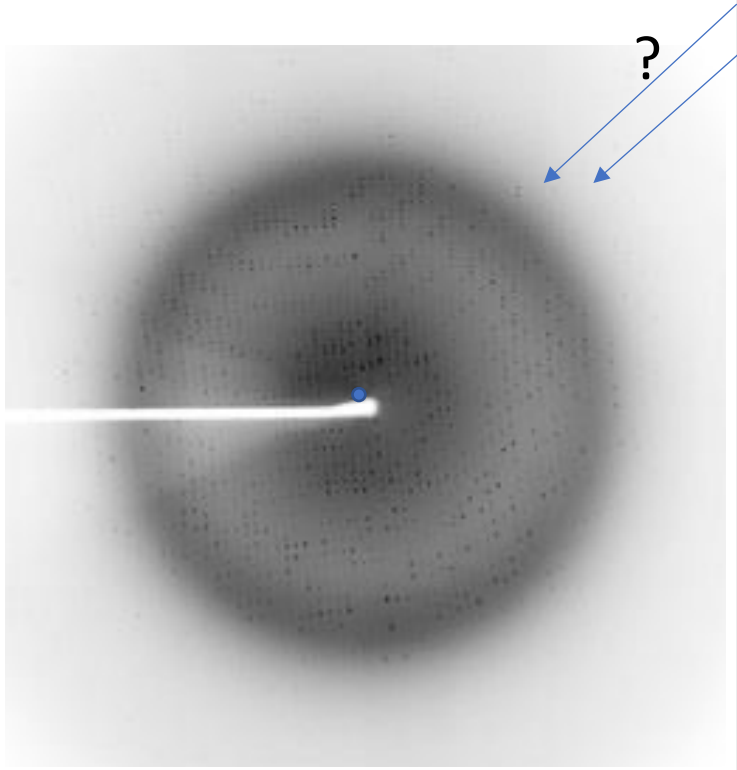
British Crystallographic Association Annual Conference held at the University of Leeds just before Easter 2022 and the ECM33 in Versailles:-

We have given clear guidelines for the ‘proper’ diffraction resolution limit, linked to the best protein model, based on the method of Diederichs and Karplus implemented as described here IUCrJ (2020) 7, 681-692.

We have emphasised the importance of the elimination, or if necessary description, of residual difference Fourier peaks which show up mismatches between the protein model and the diffraction data.

A glimpse of the variation of X-ray diffraction resolution limit choice involving the commonly used metrics in macromolecular crystallography:-

PDBj: 7ccy



Resolution cut off estimates:-

resolution of all data : 1.913
based on CC(1/2) >= 0.33 : 1.946
based on mean(I/sigma) >=2.0 : 3.037
based on R-merge < 0.5 : 2.411
based on R-meas < 0.5 : 2.497
based on completenes >=90% : 2.335
based on completeness >=50% : 2.155

Via the Diederichs and Karplus method, using the XRDa entry the resolution limit should be 2.29Å.
The depositor, Sato et al (Biochem. J. 478, 1023–1042) used 2.40 Å.

In several overview reviews of *medically important protein crystal structure studies* we have extensively tabulated the currently available protein models and diffraction data with comments on any areas of the possible improvements of their PDB files>>>

Brink, A., Jacobs, F. & Helliwell, J.R. (2022) **Trends in coordination of rhenium organometallic complexes in the Protein Data Bank** *IUCrJ 9, 180-193;*

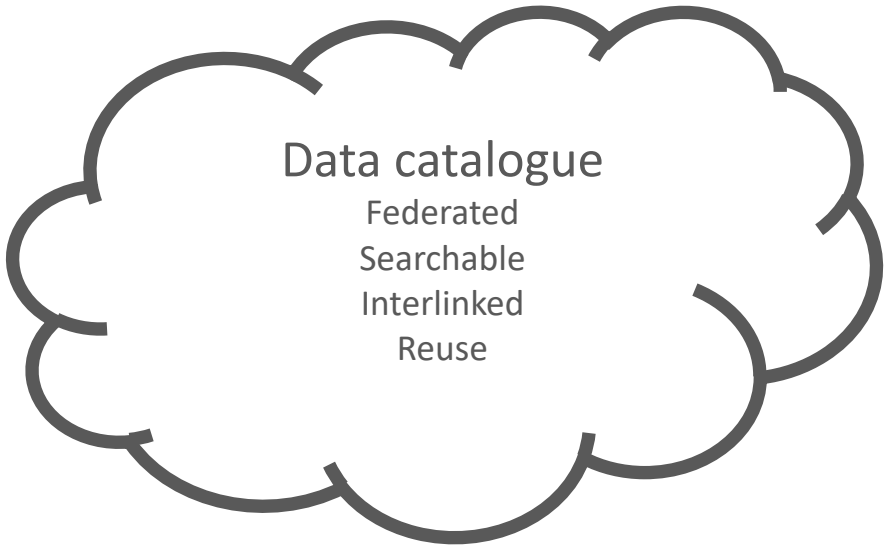
Hanau, S. & Helliwell, J.R. (2022) **6-Phosphogluconate dehydrogenase and its crystal structures** *Acta Cryst. (2022). F78, 96–112;*

Helliwell, J R (2021) **The crystal structures of the enzyme hydroxymethylbilane synthase, also known as porphobilinogen deaminase** *Acta Cryst F77, 388-398.*

A common feature of all these crystal structures are difference Fourier map peaks which have not been dealt with; 66 protein crystal structures altogether were scrutinised.

Can traditional peer review of article with data used by IUCr Journals also be applied to Facility data catalogues?

[This idea is as yet untested but CheckCif for raw data will surely help]



checkImgCIF report

```
ImgCIF checker version 2022-07-16

Checking block 5886687 in he4557img.cif

Running checks (no image download)
====================================


Testing: Required items: PASS

Testing: Data source: PASS

Testing: Axes defined: PASS

Testing: Our limitations: PASS

Testing: Detector translation: PASS

Testing: Scan range: PASS

Testing: All frames present: PASS
   All frames present and correct for SCAN1

Testing: Detector surface axes used properly: PASS

Testing: Pixel size and origin described correctly: PASS

Testing: Check calculated beam centre: PASS

Testing: Check principal axis is aligned with X: PASS
Testing presence of archive:

Testing: All archives are accessible: PASS

Running checks with downloaded images
====================================


Testing image 4: Image type and dimensions: PASS

Testing image 4: Overloaded values present: PASS

====End of Checks====
```

Conclusions

- Crystallographers have again seized opportunities to link their publications to raw diffraction data, especially Macromolecular Crystallographers
- The journal IUCrData has launched a new category of article: ***IUCr Raw Data Letters***
- New raw data sharing types of research modes have started eg ***the ESRF Paleontologists, the covid-19 Macromolecular Crystallographers….***
- All researchers can better understand the subjective choices made in their processing of raw data through to structure factors
- The collaboration with PDBj and its provision of its XRDa we see as an important development for ***definitive reusability of our MX results by biologists and medical scientists ie who are not MXers***

Acknowledgements

- Members of the IUCr DDDWG 2011-2017 and to the current Members of the IUCr Committee on Data (2017 onwards) ***https://www.iucr.org/iucr/governance/advisory-committees/committee-on-data***

- The CODATA Data Policy Committee
  - ***https://codata.org/initiatives/data-policy/international-data-policy-committee/***

- The ***checkcif for raw data Project Team***, which underpins the ***IUCrData Raw Data Letters*** initiative:
  - Loes Kroon-Batenburg (Main Editor of IUCrData's Raw Data Letters), James Hester (ANSTO and Chair of ComCIFS), Fabio Dall'Antonia, Julian Hörsch (EuroXFEL) and Andy Gotz (ESRF and PANOSC) and the staff at the IUCr Editorial Office.

GE3C

Thankyou

GE3C