Raw data opportunities for biological crystallography publishing

L.M.J. Kroon-Batenburg
Utrecht University
The Netherlands
Data publishing and management workflow

Illustration courtesy of Natalia Manova for the European OpenAIRE project
IUCr DDDWG Recommendations (top two)

• Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition of processed diffraction data (e.g. structure factor amplitudes and intensities) and coordinates, and which should obey the 'FAIR' principles, that their raw diffraction data sets should be Findable, Accessible, Interoperable and Reusable (https://www.force11.org/group/fairgroup/fairprinciples).

• A registered Digital Object Identifier (doi) should be the persistent identifier of choice (rather than a Uniform Resource Locator, url) as the most sustainable way to identify and locate a raw diffraction data set.
“IUCr Journals are now taking the lead by encouraging authors to provide a doi for their deposited original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data. “
Raw data opportunities

- What are the possibilities of raw data archiving?
- Can we adhere to the FAIR principles?
- In the refereeing process can we validate the quality and analysis of the raw data?
- Can we do new science?
FAIR for raw data in Crystallography

- Metadata schema/record
- Relevant and accurate metadata CC0-4....
- Image data formats described
- Metadata tags
- doi
# Where to archive?

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>Information on fees/costs</th>
<th>Size limits</th>
<th>Integrated with <em>Scientific Data</em>’s manuscript submission system</th>
<th>Re3data / FAIRSharing entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dryad Digital Repository</td>
<td>$120 USD for first 20 GB, and $50 USD for each additional 10 GB</td>
<td>None stated</td>
<td>Yes ✔</td>
<td>view FAIRsharing entry</td>
</tr>
<tr>
<td>figshare</td>
<td>100 GB free per <em>Scientific Data</em> manuscript. Additional fees apply for larger datasets</td>
<td>1 TB per dataset</td>
<td>Yes ✔ - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. Download instructions.</td>
<td>view FAIRsharing entry</td>
</tr>
<tr>
<td>Harvard Dataverse</td>
<td>Contact repository for datasets over 1 TB</td>
<td>2.5 GB per file, 10 GB per dataset</td>
<td>No</td>
<td>view re3data entry</td>
</tr>
<tr>
<td>Open Science Framework</td>
<td>Free of charge</td>
<td>5 GB per file, multiple files can be uploaded</td>
<td>No</td>
<td>view FAIRsharing entry</td>
</tr>
<tr>
<td>Zenodo</td>
<td>Donations towards sustainability encouraged</td>
<td>50 GB per dataset</td>
<td>No</td>
<td>view re3data entry</td>
</tr>
<tr>
<td>Mendeley Data</td>
<td>Contact repository for datasets over 10 GB</td>
<td>10 GB per dataset</td>
<td>No</td>
<td>view FAIRsharing entry</td>
</tr>
</tbody>
</table>
Findable and accessible

Data:
- OpenAire
- DataCite

Repositories of databases:
- Re3data.org
- Fairsharing.org

Discipline specific repositories:
- SBGrid
- IRRMC
- CXI

General repositories:
- Zenodo
- Figshare
- Dryad
- Research gate
- ArXiv.org
- Mendeley

Universities, National, EUDAT

Synchrotron, Neutron Facilities and XFEL:
- ESRF
- DLS
- STFC ISIS
- Store. Synchrotron
- XFELs

DOIs restricted
X-ray diffraction : 10,593

Mostly Zenodo and Figshare

<table>
<thead>
<tr>
<th>Publications (78,465)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Data (10,593)</td>
</tr>
<tr>
<td>Software (31)</td>
</tr>
<tr>
<td>Other Research Products (2,594)</td>
</tr>
<tr>
<td>Projects (587)</td>
</tr>
<tr>
<td>Content Providers (0)</td>
</tr>
<tr>
<td>Organizations (124)</td>
</tr>
</tbody>
</table>

**Transthretin Eiger 9M X-ray diffraction dataset**

**Publisher**: Figshare

<p>Transthretin X-ray diffraction dataset collected during commissioning of Eiger 9M detector on Proxima2A beamline, Synchrotron SOLEIL, France.</p>

Not SBGrid and IRRMC
DataCite: “x-ray diffraction” 15931 works

➡ Raw images data, powder data, processed data or papers

Raw data mostly:
• SBGrid
• IRRMC
• Zenodo
• CXI
• Ceon RepOD

• Figshare
• Dryad
• Mendeley
• DataShare Edinburgh
• Universities of Manchester, Leeds, Bath, Aberdeen, Cambridge, Strathclyde, Bristol, Cardiff, Utah
• Geological data
The International Union of Crystallography has for many years been advocating archiving of raw data to accompany structural papers. Recently, it initiated the formation of the Diffraction Data Deposition Working Group with the aim of developing standards for the representation of these data. A means of studying this issue is to submit exemplar publications with associated raw data and metadata. A recent study on the effects of dimethyl sulfoxide on the binding of cisplatin and carboplatin to histidine in 11 different lysozyme crystals from two diffractometers led to an investigation of the possible effects of the equipment and X-ray diffraction data processing software on the calculated occupancies and B factors of the bound Pt compounds. 35.3 Gb of data were transferred from Manchester to Utrecht to be processed with EVAL. A systematic comparison shows that the largest differences in the occupancies and B factors of the bound Pt compounds are due to the software, but the equipment also has a noticeable effect. A detailed description of and discussion on the availability of metadata is given. By making these raw diffraction data sets available via a local depository, it is possible for the diffraction community to make their own evaluation as they may wish.
“X-ray diffraction”: 15931
Restriction “Dataset”: 8804

“X-ray diffraction” AND Tanley: 15
Restriction “Dataset”: none

“X-ray diffraction” AND “crystal”: no Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

(“X-ray-diffraction" OR "raw data" OR "diffraction images") AND (macromolecule OR protein): 446

no Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

(“X-ray-diffraction" OR "raw data" OR "diffraction images") : 34568

yes Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

Helliwell: yes Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell
“X-ray diffraction”: 1292
+ Dataset: 109
Mostly Macromolecular crystallography raw data

“X-ray diffraction” AND Helliwell: none
Helliwell: 18
X-Ray Diffraction data from LapD output domain in complex with LapG, source of 4U65 structure

Data DOI: 10.15785/SBGRID/94 | ID: 94
Publication DOI: 10.7554/eLife.03650
4U65 Coordinates: Viewer, PDB (RCSB) (PDBe), MMDB
Sondermann Laboratory, Cornell University
Release Date: May 19, 2015

526 datasets

Sufficient/Valid Metadata?
Mosflm, XDS, Dials via xia2
Integrated Resource for Reproducibility in Macromolecular Crystallography

Currently indexed projects: **5265**
Currently indexed datasets: **8552**

Data downloaded from IRRMC may be freely used under the Creative Commons license CC0 (Public Domain Dedication Waiver). IRRMC strongly urges users who download data to credit the source data by using the DOI in any publications and/or derived data that make use of the downloaded data.

Diffraction project datasets IDP01325_3lus

Method: Molecular Replacement
Resolution: 1.96 Å
Space group: P 21 21 21

- Download all images (1.1 GB)
- PDB website for 3LUS
- doi:10.18430/M33LUS
**Project details**

<table>
<thead>
<tr>
<th>Title</th>
<th>Crystal structure of a putative organic hydroperoxide resistance protein with molecule of captopril bound in one of the active sites from Vibrio cholerae O1 biovar eltor str. N16961</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Nocek, B., Maltseva, N., Makowska-Grzyska, N., Kwon, K., Anderson, W., Joachimiak, A., NIAID</td>
</tr>
<tr>
<td>$R / R_{free}$</td>
<td>0.17 / 0.23</td>
</tr>
<tr>
<td>Unit cell edges [Å]</td>
<td>38.20 x 76.20 x 79.40</td>
</tr>
<tr>
<td>Unit cell angles [°]</td>
<td>90.0, 90.0, 90.0</td>
</tr>
</tbody>
</table>

**Dataset 1325-capto-x1.####.img details**

<table>
<thead>
<tr>
<th>Number of frames</th>
<th>180 (1 - 180)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance [mm]</td>
<td>292.1</td>
</tr>
<tr>
<td>Oscillation width [°]</td>
<td>1.00</td>
</tr>
<tr>
<td>Omega [°]</td>
<td>-120.0</td>
</tr>
<tr>
<td>Wavelength [Å]</td>
<td>0.97929</td>
</tr>
<tr>
<td>Experiment Date</td>
<td>2009-11-21</td>
</tr>
<tr>
<td>Equipment</td>
<td>19-ID at APS (Advanced Photon Source)</td>
</tr>
</tbody>
</table>
Raw data link in PDBe
Interoperable

**Image data formats:**
Mar345, MarCCD, ADSC, Raxis, Oxford, CMOS RDI, Pilatus (imgCIF/cbf), Eiger (HDF5) ....CSPAD,AGIPD...

Software packages can deal with most image formats:
HKL3000/XDS/d*Trek/Mosflm/Dials/EVAL

**Vocabulary: metadata tags:**
Plethora of Ascii key-words, imgCIF, Nexus
<table>
<thead>
<tr>
<th>Minimal Metadata</th>
<th>imgCIF tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data binary format</td>
<td>• _array_structure_byte_order, _array_structure_compression_type</td>
</tr>
<tr>
<td>• Number of pixels, pixel size (binning mode)</td>
<td>• _array_structure_list.index;</td>
</tr>
<tr>
<td>• Beam Center (mm, pixels)</td>
<td>• _array_structure_list.dimensions</td>
</tr>
<tr>
<td>• Origin of data frame</td>
<td>• _array_element_size.size</td>
</tr>
<tr>
<td>• Wavelength</td>
<td>• _diffrn_detector_element.center[1]</td>
</tr>
<tr>
<td>• Rotation axis</td>
<td>• _diffrn_detector_element.center[2]</td>
</tr>
<tr>
<td>• Rotation range per frame</td>
<td>• _diffraction_radiation.wavelength.wavelength</td>
</tr>
<tr>
<td>• Axes and offsets</td>
<td>• _diffrn_scan_axis.axis_id,</td>
</tr>
<tr>
<td>• Detector-to-sample distance</td>
<td>• _diffrn_scan_axis.displacement_start</td>
</tr>
<tr>
<td></td>
<td>• _diffrn_scan_axis.displacement.increment</td>
</tr>
<tr>
<td></td>
<td>• _axis.id, _axis.vector[1].., _axis.offset[1]..</td>
</tr>
</tbody>
</table>
Implicitly assumed (Expert knowledge)
- Orientation of rotation axis
- Rotation direction
- Detector swing angle (0°)
- Polarization
- Detector type

Advanced
- Sensor thickness
- Baseline offset
- Overflow level
- Polarization
- Gain
- Detector swing
- Multi axis goniometer
- Exposure time
- Bad pixels
- Time stamp

COMCIFS/CommDat: imgCIF metadata and CheckCIF
HDRMX and Nexus

continued...
Raw data re-use

Reasons for reprocessing:
- Multiple lattices: % overlap (if we can go to CC1/2 0.14 this should matter)
- TDS/background (not solved in integration; also streaks not accounted for)
- Resolution cut-off
- Unsolved structure
- Diffuse scattering (packing disorder or internal mobility)
- Incommensurate modulation
Monoclinic: 42.6 75.2 76.5 90 106.2 90 -> C-centered orthorhombic 42.6 146.4 75.2 90 90 90

Dirax finds two matrices, Data processed with EVAL

Problems with reprocessing:
1) offset of omega-axis (~1°)
2) twin 177.6° rotation around (0,1,-2) in orthorhombic lattice

23% deconvoluted 15% overlapping

Beamline: 4.2.2 ALS
Detector: CMOS_8M RDI
Format: img
Header: ADSC
**BJP1_3m8t**

<table>
<thead>
<tr>
<th>3M8T</th>
<th>2008</th>
<th>Resolution 1.33 Å</th>
<th>MOSFLM</th>
</tr>
</thead>
</table>

| unitcell: 47.2 45.2 77.4 100.9 89.8 118.1 |

Symmetry P1, two independent molecules

Second fragment 3.8° rotation, arbitrary axis
5% overlap with first matrix

![Image of EVAL15 box](image1.png)
Multiple crystal forms

E. coli enzyme N-acetylneuraminic lyase

Pathological macromolecular crystallographic data affected by twinning, partial-disorder and exhibiting multiple lattices for testing of data processing and refinement tools


Data Records
The datasets (raw diffraction images) discussed in this manuscript have been deposited in the publicly available database zenodo at, https://doi.org/10.5281/zenodo.54568 and 10.5281/zenodo.1240503. Structural models and processed structure factor data deposited in the PDB are available under the accession codes given in Table 1, with the exception of dataset Y137A, as the R factor indices were not satisfactory for PDB deposition.
Four crystal form I structures have incommensurate modulation q-vector: ~0.16 0.0 ~0.43 and twinning (-h,-k,h+l)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Res</th>
<th>Crystal form and cell parameters</th>
<th>Obliquity* (ω)</th>
<th>Twinning fraction*</th>
<th>Twin on</th>
<th>Twin off</th>
<th>PDB code</th>
<th>Diamond Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type apo</td>
<td>2.30 Å</td>
<td>P2₁, crystal form I a = 54.8 b = 142.2 c = 84.2 α = 90.00 β = 108.97 γ = 90.00</td>
<td>0.019</td>
<td>0.372</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.200</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.267</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.251</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.319</td>
</tr>
<tr>
<td>Wild type pyruvate complex</td>
<td>1.65 Å</td>
<td>P2₁, crystal form I a = 54.7 b = 142.5 c = 83.6 α = 90.00 β = 109.16 γ = 90.00</td>
<td>0.070</td>
<td>0.334</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.201</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.245</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.256</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.293</td>
</tr>
<tr>
<td>E192N apo</td>
<td>1.80 Å</td>
<td>P2₁, crystal form I a = 54.6 b = 142.8 c = 84.3 α = 90.00 β = 108.8 γ = 90.00</td>
<td>0.130</td>
<td>0.463</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.195</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.244</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.272</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.320</td>
</tr>
<tr>
<td>E192N pyruvate complex</td>
<td>1.80 Å</td>
<td>P2₁, crystal form I a = 56.9 b = 143.0 c = 83.9 α = 90.00 β = 109.8 γ = 90.00</td>
<td>0.000</td>
<td>—</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.178</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.209</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.187</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.223</td>
</tr>
<tr>
<td>E192N + pyruvate + THB**</td>
<td>2.05 Å</td>
<td>P2₁, crystal form I a = 57.0 b = 143.7 c = 84.3 α = 90.00 β = 109.9 γ = 90.00</td>
<td>0.130</td>
<td>—</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.192</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.238</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.191</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.242</td>
</tr>
<tr>
<td>Y137A pyruvate complex</td>
<td>1.80 Å</td>
<td>P2₁, crystal form I a = 54.7 b = 142.2 c = 83.6 α = 90.00 β = 109.9 γ = 90.00</td>
<td>0.119</td>
<td>0.149</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.287</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.331</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.296</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.357</td>
</tr>
<tr>
<td>Y137A pyruvate, ManNac and Neu5Ac complex</td>
<td>2.00 Å</td>
<td>P2₁, crystal form I a = 56.1 b = 143.5 c = 83.6 α = 90.0 β = 109.6 γ = 90.00</td>
<td>0.094</td>
<td>0.497</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.183</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.236</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.265</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.321</td>
</tr>
<tr>
<td>Wild type apo</td>
<td>1.90 Å</td>
<td>P2₁, crystal form II a = 84.3 b = 95.9 c = 91.4 α = 90.00 β = 115.33 γ = 90.00</td>
<td>2.10</td>
<td>—</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.198</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.226</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.197</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.225</td>
</tr>
<tr>
<td>E192N/Y137F pyruvate complex</td>
<td>1.80 Å</td>
<td>P2₁, crystal form III a = 78.0 b = 116.7 c = 83.7 α = 90.0 β = 118.06 γ = 90.00</td>
<td>0.290</td>
<td>0.328</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.156</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.183</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.206</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.228</td>
</tr>
<tr>
<td>E192N + pyruvate complex</td>
<td>1.85 Å</td>
<td>P2₁, crystal form III a = 78.1 b = 116.5 c = 83.7 α = 90.00 β = 116.5 γ = 90.00</td>
<td>0.150</td>
<td>0.096</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.165</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.186</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.174</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.193</td>
</tr>
<tr>
<td>E192N + pyruvate</td>
<td>1.45 Å</td>
<td>P2₁, crystal form IV a = 78.3 b = 108.6 c = 141.8 α = β = γ = 90.00</td>
<td>0.000</td>
<td>—</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.191</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.201</td>
<td>R&lt;sub&gt;factor&lt;/sub&gt; = 0.188</td>
<td>R&lt;sub&gt;free&lt;/sub&gt; = 0.205</td>
</tr>
</tbody>
</table>
Incommensurate modulation

E. coli enzyme N-acetyl-neuraminic lyase
Validation

Validation of the structural model:

- IUCr CheckCIF
- wwPDB validation report

Structural model in the light of the processed data (hkl, I, σ(I))

“CheckCIF” for raw images:

- Check for core metadata
- Validation of the analysis of raw data:
  - is everything in the diffraction pattern understood?
  - Referee or automatic validation?
Automatic validation of diffraction image analysis?

Thaumatin data from I03 Diamond Light Source  Detector ADSC 3000x3000 pixels

\[ \rho_{\text{obs}} \]

\[ \rho_{\text{model}} = JP_i + \sum_m M P_{im} + ax_i + by_i + c \]

Reconstruction from modelling 5 images
Modelled image with EVAL

Iobs

Imodel
Distribution of pixel values?

\[ I_{\text{obs}} \]

\[ I_{\text{bg}} = ax_i + by_i + c \]

The diffuse background reconstructed from observations per box
Conclusions

• FAIR?

FAIR

- Metadata schema/record
- Relevant and accurate metadata CC0-4....
- Image data formats described
- doi
- Metadata tags

• COMCIFS/CommDat: imgCIF metadata and CheckCIF
• HDRMX and Nexus
• Referees can have a look at the images and the indexing
• Automatic validation of raw data interpretation?