Do we treat Macromolecular Crystallography data FAIR?

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 Re-usable (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original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data. “
FAIR

• **Findable:** easy to identify and find for both humans and computers, with metadata that facilitate searching for specific datasets,

• **Accessible:** stored for long term so that they can easily be accessed and/or downloaded with well-defined access conditions, whether at the level of metadata, or at the level of the actual data,

• **Interoperable:** ready to be combined with other datasets by humans or computers, without ambiguities in the meanings of terms and values,

• **Reusable:** ready to be used for future research and to be further processed using computational methods. This requires adequate information about how the data were obtained and processed (provenance) and an appropriate license.
FAIR for raw data in MX

FAIR

- Metadata schema/record
  - Doi
  - Protocol
  - userID

- Relevant and accurate metadata CC0-4....

- Image data formats described
  - Metadata tags
Raw data in MX

• What are the possibilities of raw data archiving?
• Can we adhere to the FAIR principles?
• Do we reuse the data?
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

Data Policies
DataCite: “x-ray diffraction” 19488 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
X-Ray Diffraction data from SARS-CoV-2 Nucleocapsid N2b domain, source of 6WZO structure
Qiaozhen Ye
X Ray Diffraction published via SBGrid Data Bank
Native dataset for SARS-CoV-2 Nucleocapsid (N) dimerization domain, P1 form

No citations were reported. No usage information was reported.

https://doi.org/10.15785/sbgrid/785
Findable?

diffraction AND "SARS-CoV-2" NOT zenodo NOT SBGrid → 24

SARS-CoV-2 → 29
COVID-19 → 13

Metadata!
→ “X-ray diffraction”: 1462 + Dataset: 117  Mostly Macromolecular crystallography raw data

→ Community: macromolecules AND diffraction: 151  
→ Diffraction AND Covid-19: 78 entries
→ Diffraction AND “SARS-Cov-2”: 79 entries
→ Diffraction AND protein: 79 entries

Metadata:
protein sample reference to pdb

Metadata:
Description of experiment image headers or Nexus/HDF5
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

- Reprocessing Instructions
  beam center x=99.3, y=100.0 indexed in P21 with HKL2000. Distance is 217 (not 220 as indicated in the header)

- Reprocessing Data
  xia2-2d  xia2-3d  xia2-3dii  xia2-dials

Sufficient/Valid Metadata?
Mosflm, XDS, Dials via xia2
Currently indexed projects: 5477
Currently indexed datasets: 8839

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.
### Project details

<table>
<thead>
<tr>
<th><strong>Title</strong></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><strong>Authors</strong></td>
<td>Nocek, B., Maltseva, N., Makowska-Grzyska, N., Kwon, K., Anderson, W., Joachimiak, A., NIAID</td>
</tr>
<tr>
<td><strong>R / R&lt;sub&gt;free&lt;/sub&gt;</strong></td>
<td>0.17 / 0.23</td>
</tr>
<tr>
<td><strong>Unit cell edges [Å]</strong></td>
<td>38.20 x 76.20 x 79.40</td>
</tr>
<tr>
<td><strong>Unit cell angles [°]</strong></td>
<td>90.0, 90.0, 90.0</td>
</tr>
</tbody>
</table>

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

| **Number of frames** | 180 (1 - 180) |
| **Distance [mm]** | 292.1 |
| **Oscillation width [°]** | 1.00 |
| **Omega [°]** | -120.0 |
| **Wavelength [Å]** | 0.97929 |
| **Experiment Date** | 2009-11-21 |
| **Equipment** | 19-ID at APS (Advanced Photon Source) |
John Berrisford: out of the 9665 X-ray entries that were released in 2019 we have DOI's for raw images in 205 of these entries.
An unprecedented number of scientific efforts are taking place worldwide in order to help combat the new coronavirus epidemic (COVID-19). One of the biggest challenges in this fast-moving situation is to share data and findings efficiently. PDBe-KB COVID-19 Data Portal provides a platform for researchers to upload, share, and access relevant data and resources.

**PDBe-KB COVID-19 Data Portal**

Spike protein S1 attaches the virion to the cell membrane by interacting with host receptor, initiating the infection. Binding to human ACE2 and CLEC4M/DC-SIGN receptors and internalization of the virus into the endosomes of the host cell induces conformational changes in the S glycoprotein. Proteolysis by cathepsin CTSL may unmask the fusion peptide of S2 and activate membranes fusion within endosomes.

Spike protein S2 mediates fusion of the virion and cellular membranes by acting as a class I viral fusion protein. Under the current model, the protein has at least three conformational states: pre-fusion native state, pre-hairpin intermediate state, and post-fusion hairpin state. During viral and target cell membrane fusion, the coiled coil regions...
PDBe-KB ➤ Spike glycoprotein

Gene: S

Organism: Severe acute respiratory syndrome coronavirus 2 (2019-nCoV)

Uniprot: P0DTC2 [go to UniProt]

Biological function: May down-regulate host tetherin (BST2) by lysosomal degradation, thereby counteracting its antiviral activity [go to UniProt]
Whenever necessary and possible, we intend to start our re-analysis from reprocessing of the original diffraction images [28]. However, quite often we were unable to obtain the diffraction data despite the IUCr recommendation [29] and an earnest appeal from the community to make diffraction data related to CoV-2 public.
Reuse of processed data

Andrea Thorn

Re-refinement: Isolde
Validation: Auspex, Molprobity
Links to: PDB-Redo and Buster (Global Phasing)

→ 83 links to raw diffraction data:
  74 Zenodo: main protease (PANDDA project)
  9 IRRMC
  (2 in SBGrid are not mentioned)
(Re)processing available raw image data - Take-1

<table>
<thead>
<tr>
<th>20200421</th>
<th>Looking at SARS-CoV-2 papain-like protease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Looking at SARS-CoV-2 NSP16/NSP10 structures</td>
</tr>
<tr>
<td></td>
<td>Looking at SARS-CoV-2 NSP3 structures</td>
</tr>
<tr>
<td></td>
<td>Looking at SARS-CoV-2 NSP15 Endoribonuclease</td>
</tr>
<tr>
<td></td>
<td>Looking at SARS-CoV-2 Nsp9 RNA binding protein</td>
</tr>
</tbody>
</table>

In total 10 data sets (from proteindiffraction.org) are being re-processed.
Re-refinement of the processed data is ongoing.
**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
Reasons for reprocessing:
• Multiple lattices: % overlap (if we can go to $CC_{1/2}$ 0.14 this should matter)
• TDS/background (not solved in integration; also streaks not accounted for)
• Resolution cut-off
• Anisotropic data
• Unsolved structure
• Diffuse scattering (packing disorder or internal mobility)
• Incommensurate modulation
Conclusions

- FAIR?
  - Metadata schema/record
  - Doi
  - Protocols
  - userID
  - Relevant and accurate metadata
  - CC0-4....
  - Image data formats described
  - Metadata tags

- IUCr: imgCIF dictionary
- HDRMX NXmx Gold Standard using Nexus/HDF5