How to make structures better in the future and the use of published raw data in crystallographic software development

L.M.J. Kroon-Batenburg
Utrecht University
The Netherlands
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.
FAIR for raw data in crystallography

FAIR

Metadata schema/record

Doi
Protocol
userID

Image data formats described
Metadata tags

Relevant and accurate metadata
CC0-4....
DataCite: “x-ray diffraction” 47389 works

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

**Raw data mostly:**
- SBGrid
- IRRMC
- Zenodo
- CXI
- MX-RDR/RepOD

- Figshare
- Dryad
- Mendeley
- DataShare Edinburgh
- Universities of Manchester, Leeds, Bath, Aberdeen, Cambridge, St. Andrews, Strathclyde, Bristol, Cardiff, Utah
- Geological data

provides persistent identifiers (DOIs) for research data and other research outputs
Raw data archive options for Chemical Crystallography

- Typical data set size (dependent on resolution and detector type): 100 MB – 4 GB

- Do (service) crystallographers have the infrastructure to archive all raw data?
  - University repository
  - Public repository
  - Neutron/Synchrotron facility
  - Hard drive/DVD/tape

- Funding agency adhere to Open Science policies:
  - all research data should become publicly available
Raw data archive options for Chemical Crystallography

- Typical data set size (dependent on resolution and detector type): 100 MB – 4 GB

- Do (service) crystallographers have the infrastructure to archive all raw data?
  - University repository
  - Public repository -> Zenodo
  - Neutron/Synchrotron facility
  - Hard drive/DVD/tape

- Funding agency adhere to Open Science policies:
  - all research data should become publicly available
Which raw data to archive?

- Data related to high-impact, complicated or remarkable structures with demanding structure solution and refinement
- Data with particular features, like (non-)merohedral twinning, multiple lattices, diffuse scattering, incommensurate modulation
- Data of from which the structure could not be solved
Where to archive?

Zenodo

Featured communities

Chicago COVID-19 Response
This repository community collects research outputs and efforts in Chicago. Users are encouraged to upload their re-discovery of information. Although Open Access articles are
Curated by: saragon

Recent uploads

Natural history specimens collected and/or identified and deposited.


Need help?

Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:
- Uploading your research data, software, preprints, etc.
Communities created and curated by Zenodo users

crystallography|

Showing 0 to 10 out of 8309 communities.

Coronavirus Disease Research Community - COVID-19

This community collects research outputs that may be relevant to the Coronavirus Disease (COVID-19) or the SARS-CoV-2. Scientists are encouraged to upload their outcome in this collection to facilitate sharing and discovery of information. Although Open Access articles and datasets are recommended, also closed and restricted access material are accepted. All types of research outputs can be included in this Community (Publication, Poster, Presentation, Dataset, Image, Video/Audio, Software, Lesson, Other).

Want your own community?
It's easy. Just click the button to get started.

- Curate — accept/reject what goes in your community collection.
- Export — your community collection is automatically exported via OAIPMH
- Upload — get custom upload link to send to people

Biodiversity Literature Repository

European Commission Funded Research (OpenAIRE)
Macromolecular Crystallography

Macromolecular crystallography community: people who use X-ray diffraction to gain insight into the structure and function of biological macromolecules such as proteins, DNA, carbohydrates & similar.

Curated by: graeme

Structural Genomics Consortium - Diamond Light Source I04-1 XChem Fragment screening by X-ray crystallography

Fragment-based screening is now well-established as a powerful approach to early drug, or 'lead', discovery. The principle is to identify weakly-binding compounds ('fragments') by screening a limited library of compounds, with resulting hits serving...

Curated by: SGC-DLS-XCHEM-I04-1-FS

Fragment based drug discovery by X-ray crystallography

Fragment based drug discovery (FBDD) has become an increasingly important tool for finding hit compounds for difficult targets. The technique utilises smaller than drug-like compounds to identify low potency, high quality leads. Libraries containing...

Curated by: graeme

Chemical Crystallography

X-ray diffraction data for chemical crystallography, including but not limited to data from I19-1 and I19-2 at Diamond Light Source

Curated by: graeme

Want your own community?

It's easy. Just click the button to get started.

- Curate — accept/reject what goes in your community collection.
- Export — your community collection is automatically exported via OAI-PMH
- Upload — get custom upload link to send to people
Recent uploads

Diffraction data for CCDC 2031535–2031545 and 2031863–2031890. Claudetite I at various temperatures and pressures

Gurka, Piotr A.

Diffraction data for claudetite I, monoclinic arsenic(III) oxide polymorph collected at room and low temperature (single-crystal diffraction data corresponding to CCDC 2031535-2031545) and at high pressure (powder diffraction data corresponding to CCDC 2031863-2031890).

Uploaded on December 14, 2020
Published in CrystEngComm, vol. 23, issue 3, pp. 638-644.

Diffraction data for CCDC 1949980–1949982

Gurka, Piotr A.

Diffraction data for: 10-hexyl-1-(10H-phenoxazin-10-yl)acridin-9(10H)-one 10-hexyl-2-(10H-phenoxazin-10-yl)acridin-9(10H)-one 10-hexyl-3-(10H-phenoxazin-10-yl)acridin-9(10H)-one

Uploaded on November 13, 2020
CSD Entry: JALNAO

Space Group: P 2/c (14), Cell: a 12.679(5) Å, b 8.8317(3) Å, c 15.5666(6) Å, α 90° β 109.439(4)° γ 90°

 Additional details

Deposition Number 1515700
Data Citation Piotr Kurzep, Łukasz Skórka, Małgorzata Ządurska, Piotr A. Głurka, Marzena Banasiak, Bolesław Kozanińczyk, Irena Kuliszewicz-Bojar CCDC 1515700: Experimental Crystal Structure Determination, 2017, DOI: 10.5617/ccdc.csd.c.1515700
Deposited on 06/11/2016

Raw data DOI(s)
DOI 10.5281/zenodo.4271549
Carefully look at your data: Diaquobis(salicylato)copper(II)

- Reflections can be indexed in an orthorhombic or monoclinic lattice
- Structures could be solved for $h$=even

### Table 3. Experimental details for the orthorhombic setting.
Refined as an inversion twin. Contributions of the monoclinic structure removed from the structure factors.

<table>
<thead>
<tr>
<th>Molecular formula</th>
<th>CuC$_6$H$_4$O$_3$(H$_2$O)$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula weight</td>
<td>373.79</td>
</tr>
<tr>
<td>Crystal system</td>
<td>Orthorhombic</td>
</tr>
<tr>
<td>Space group</td>
<td>Pca$_2$</td>
</tr>
<tr>
<td>$a$ / Å</td>
<td>7.6538(11)</td>
</tr>
<tr>
<td>$b$ / Å</td>
<td>11.7378(3)</td>
</tr>
<tr>
<td>$c$ / Å</td>
<td>31.5707(5)</td>
</tr>
</tbody>
</table>

### Table 4. Experimental details for the monoclinic setting.
Refined as a rotation twin. Contributions of the orthorhombic structure removed from the structure factors.

<table>
<thead>
<tr>
<th>Molecular formula</th>
<th>Cu(C$_6$H$_4$O$_3$)(H$_2$O)$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula weight</td>
<td>373.79</td>
</tr>
<tr>
<td>Crystal system</td>
<td>Monoclinic</td>
</tr>
<tr>
<td>Space group</td>
<td>$P2_1/a$</td>
</tr>
<tr>
<td>$a$ / Å</td>
<td>7.6538(2)</td>
</tr>
<tr>
<td>$b$ / Å</td>
<td>11.7408(3)</td>
</tr>
<tr>
<td>$c$ / Å</td>
<td>31.8088(12)</td>
</tr>
<tr>
<td>$\beta$ / °</td>
<td>96.9831(2)</td>
</tr>
</tbody>
</table>
Reciprocal space

(h0l)  (h1l)  (1kl)

Ortho

Mono

Random stacking faults

On the blue stacking lattice the structure is ordered.

Raw data would complement the extensive archives of derived (coordinates) and processed (structure factors) data

- CheckCIF: syntax, consistency, completeness, validation
  Acta Cryst. C requires unmerged structure factors
- Macromolecular crystallography: PDB validation report wwPDB, Structure factors CIF and check with coordinates

- CheckCIF for raw data (CommDat/COMCIFS): *three levels*
  ✓ Check for consistency, completeness of metadata
  ✓ Reprocessing: sufficient metadata?
  ✓ How much of the data is understood?
  indexing; lattice symmetry; pseudo merohedral twinning; incommensurate; data reduction; diffuse scattering
Single-crystal X-ray diffractometry data for a sample of NiCl₂-dppe collected on beamline I19-2 at Diamond Light Source

Warren, Mark R; Allan, David R

Data curator(s)
Williams, Benjamin Heathcote

Single-crystal X-ray diffractometry data for a sample of [1.2 Bis(diphenylphosphino)ethane] dichloronickel(II) (NiCl₂-dppe, [Cl₂Ni(C₂H₄PO)₂(C₆H₅)₂]Cl₂) was collected at Diamond Light Source I19-2 on 2015-05-18, and it is publicly available for users to test data reduction routines.

The sample was prepared as follows:
Nickel chloride (II) hexahydrate (1 g, 2 mmol) was heated under vacuum to produce anhydrous nickel chloride (II) with a visible colour change from green to yellow. The resulting solid was taken up in ethanol (5 ml) and added to 1.2-bis(diphenylphosphino)ethane (dppe) (0.837 g, 2 mmol) in ethanol (10 ml). The solution was refluxed for 3 hour after which the solvent was evaporated. The small red crystals were purified by recrystallisation in acetone (70% yield).

The sample was held at an approximate temperature of 150 K and the illuminating beam had a wavelength of 0.6889 Å (17.097 keV). The detector was held at 2θ = 25° throughout.

Inventory of data:
010_Ni_dppe_CL2_150K01 — 138° ω scan, 0.4° images, 0.4° steps, 325 images; k = 45°, q = 160°.
010_Ni_dppe_CL2_150K02 — 138° ω scan, 0.4° images, 0.4° steps, 325 images; k = 45°, q = 220°.
010_Ni_dppe_CL2_150K03 — 138° ω scan, 0.4° images, 0.4° steps, 325 images; k = 45°, q = 80°.
010_Ni_dppe_CL2_150K04 — 198° ω scan, 0.4° images, 0.4° steps, 495 images; k = 0°, q = 80°.

Files (152 MB)

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>010_Ni_dppe_CL2_150K01.tar.gz</td>
<td>69.6 MB</td>
</tr>
<tr>
<td>010_Ni_dppe_CL2_150K02.tar.gz</td>
<td>69.7 MB</td>
</tr>
<tr>
<td>010_Ni_dppe_CL2_150K03.tar.gz</td>
<td>69.7 MB</td>
</tr>
<tr>
<td>010_Ni_dppe_CL2_150K04.tar.gz</td>
<td>106.2 MB</td>
</tr>
</tbody>
</table>
### CBF: VERSION 1.5, CBFlib v0.7.8 - PILATUS detectors

data_010_Ni_dppe_C1_2_158K01_00001

_array.data.header_convention "PILATUS_1.2"

_array.data.header_contents

# Detector: PILATUS 300K, S/N 3-0104, Diamond
# 2015-05-10T21:41:55.427
# Pixel size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.000320 m
# Exposure time 3.0700008 s
# Exposure period 0.4000000 s
# Tau = 199.1e-09 s
# Count cutoff 999974 counts
# Threshold setting: 16000 ev
# Gain setting: automat (vrf = 1.000)
# N_excluded_pixels = 31
# Excluded pixels: badpixel mask.tif
# Flatfield: FF_p3_0104_E20000_T1000_20150711_v3.tif
# Trim file: p3_0104_E20000_T10000.bin
# Image path: ...

### miniCBF

### Full ImgCIF/CBF
How “Raw” can/should the data be?

BrukerNonius KappaCCD images
FR591 rotating anode

-> Distorted/warped image data

Data from in-house diffractometer with commercial software package
Shouldn’t the images be unwarped, corrected for sensitivity and dark current and dezingered to be re-usable?
Diffraction data for CCDC 1515700

Gurka, Piotr A
Project leader(s)
Kuliszewska-Bajer, Irena

Diffraction data for 8.16-dioctyl-8,16-dihydroacridino[2.1-a]acridine-5,13-dione

Polish National Science Centre, Grant No. 2015/17/B/ST5/00179

Indexed in OpenAIRE

EVAL software
Images formats and compressions should be publically available
Vendors should make provisions for exporting raw data in a re-usable format.
IUCrData, the peer-reviewed open-access data publication from the International Union of Crystallography (IUCr), is launching a new section for authors to describe their unprocessed or "raw" diffraction images. This is a collaborative innovation of IUCr Journals with the IUCr Committee on Data.

The new section will publish short descriptions of crystallographic raw data sets in the biological, chemical or materials science fields and provide a persistent link to the location of the raw data. This will allow researchers to attract attention to particular features of the data that could be of interest to methods and software developers or may be relevant to the structural interpretation.

The IUCr will adhere to the FAIR principles for which the correctness and completeness of the metadata are crucial, and these will be central to the reviewing process.
The new section will be accepting submissions from the autumn and anyone wanting to know more should contact the IUCr Editorial Office (med@iucr.org).
CheckCif for raw data

- Parser to extract meta data and write it in imgCIF
- Check validity of meta data
Conclusions

• It is strongly encouraged to archive raw data sets associated with publications
• It is possible to archive raw data sets for free: Zenodo
• The CX community has started publishing raw data
• The CX community should discuss what is meant with “Raw” data: unprocessed vs. re-usable