CURRENT PRACTICES IN (HOME LAB) RAW DATA ARCHIVAL

Responses to a survey published at:

A SNAPSHOOT OF CURRENT ARCHIVAL PRACTICE

Mainly ‘Do It Yourself’ with resources readily to hand in the lab

Around 20% use more ‘modern’ (ie networked) approaches
WHY NOT ARCHIVE RAW DATA?

38/193 people DON’T archive raw data

Half of these probably would if there was support available
Largest response is no

Majority perform some sort of activity

Very little active review over time
Very few ‘advanced’ approaches

Mostly tools provided already by operating system

Requires prior knowledge to search
EXTENT OF ‘OPENNESS’

Is your raw data externally available?
186 responses

- Yes (e.g. open access; collaborators; other researchers requesting access)
- No
WHAT IF...?

If it were to become Policy to make funded research raw diffraction data accessible after 3 years since measurement, would you endeavour to comply?

183 responses

- 47.5%: Yes, I would do this for all my raw data
- 43.2%: Yes, but I would only do so in cases where I had to
- 9.3%: No
AFFORDABILITY

Would you anticipate being able to pay for external archive/repository services and facilities e.g. through research grants or institutional funding

187 responses

- Yes: 7.5%
- No: 59.9%
- Maybe: 32.6%
Over half unaware of any governing policy

Funding agencies are the most prevalent, with institutions also featuring
Complex structures arising from the self-assembly of a simple organic salt

Riccardo Montis, Luca Fusaro, Andrea Faqui, Michael B. Hursthouse, Nikolay Tumanov, Simon J. Coles, Terry L. Threlfall, Peter N. Horton, Rachid Sougrat, Anais Lafontaine, Gerard Coquerel & A. David Roe

Data availability

Full crystallographic details in CIF format have been deposited in the Cambridge Crystallographic Data Centre database (deposition numbers: phase 1, CCDC 1540139 (unconstrained) and 1540140 (constrained); phase 2, CCDC 1897427; phase 3, CCDC 1540141; phase 4, 1897428). Copies of this information may be obtained free of charge from deposit@ccdc.cam.ac.uk or http://www.ccdc.cam.ac.uk. Raw single-crystal diffraction data corresponding to the structures of phases 1–4 have been deposited in the Zenodo repository at the following locations: phase 1, https://doi.org/10.5281/zenodo.2595089; phase 2, https://doi.org/10.5281/zenodo.2585776; phase 3, https://doi.org/10.5281/zenodo.2593670; phase 4, https://doi.org/10.5281/zenodo.2593677. The CCD images (in either Rigaku IM or Bruker KCD format) have been deposited, along with instrument parameters and all files associated with image processing. This will enable the reader to fully validate these structural models and, for those who wish to investigate alternative approaches to modelling these results or develop them further, it will be possible to do so without having to synthesize the materials and collect diffraction data. NMR results are...
CONCLUSIONS?

This behaviour reflects general data management practice

‘Archival’ can imply keeping data safely out of sight, but some current data management practice could easily be converted to more ‘open’

Only a modest amount of culture change required

‘Lab finances’ are not a significantly influencing factor for local archive, but are for publishing

Some infrastructure and guidance development is necessary…

• Index/Search system (not reliant on operating system/single lab ID)
• Persistent IDs
• Standard archive approach (metadata + packaging format)
• Mechanism to clearly link raw data with ‘results’ data
• System to establish if an image is ‘clean’
• System to establish if all Bragg diffraction in raw data accounted for in the model
• Ability to have control over when/if to publish

A start: IUCrData to publish "Raw" Data Letters…
## TODAY’S AGENDA

### Day 2: Thursday 12 August
### Session 2: What would a ‘standard’ look like? – best practice

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00-14:10</td>
<td>Simon Coles (Southampton)</td>
<td>Current practices in raw data archival (home laboratories)</td>
</tr>
<tr>
<td>14:10-14:35</td>
<td>Graeme Winter (Diamond)</td>
<td>Future outlook for a curated raw data archive</td>
</tr>
<tr>
<td>14:35-15:00</td>
<td>Loes Kroon-Batenburg (Utrecht)</td>
<td>How to make structures better in the future and the use of published raw data in crystallographic software development</td>
</tr>
<tr>
<td>15:00-15:20</td>
<td>Coffee</td>
<td></td>
</tr>
<tr>
<td>15:20-15:45</td>
<td>Teodor Ivănoaica (Extreme Light Infrastructure)</td>
<td>Data Policy at Large Research Infrastructure. Data Retention challenge</td>
</tr>
<tr>
<td>15:45-16:10</td>
<td>Natalie Johnson (CCDC)</td>
<td>The value of integration of raw data with publishing procedures and results databases</td>
</tr>
<tr>
<td>16:35</td>
<td>Panel discussion</td>
<td>16:35 (10:35)</td>
</tr>
<tr>
<td>17:00</td>
<td>Close of Day 2</td>
<td></td>
</tr>
</tbody>
</table>