FUTURE OUTLOOK FOR CURATED DATA ARCHIVE
OVERVIEW

• Historical viewpoint - archiving in crystallography

• Features of a useful raw data archive

• Curation

• Supporting data archiving - cost / benefit

• Conclusions
I am considering biological and chemical crystallography as “the same problem” here.
CRYSTALLOGRAPHY AS A DATA SCIENCE

Crystallography highly data driven -

• Determine results fully from experimental data and prior knowledge

• The “shape” of the data are well known in advance (i.e. not “messy”)

• Influence of interpretation much reduced compared to e.g. geology

• Process significantly automated, powerful library of tools used for analysis

• Comparable with radio astronomy as observational / data driven science
ARCHIVING IN CRYSTALLOGRAPHY

Long history of data banks / archives in X-ray crystallography

- CSD - 1965
- PDB - 1971
- ICSD - 1979

Crystallography pioneered open data archives

Computer-based archiving fundamental
CONTENT OF AN ARCHIVE RECORD

• 3D atomic coordinates
• Report of experiment - maybe
• Reference to publication - often
• Experimental data (processed) - maybe
• Reference to external data - maybe
CONTENT OF A USEFUL ARCHIVE RECORD

- 3D atomic coordinates
- Report of experiment
- Reference to publication
- Experimental data (processed)
- Reference to external data
ACCESS TO DATA ARCHIVE

- Data are generally freely available in predefined format (CIF, PDB, mmCIF, ...)
- Services may not be freely available e.g. ability to usefully search, links to other data archives
- wwPDB made up of RCSB / EBI (PDBe) and PDBj - “competition” between databases, built on the same underlying databank
- CSD funded by CCDC as not for profit company
FEATURES OF USEFUL RAW DATA ARCHIVE

- Easy to search, well integrated with existing data archives (PDB, CSD etc.)
- Inclusive / open to all depositors / open to all users
- Curated
- Funded / sustainable / long lived
EXAMPLE - ICAT

• Designed for STFC facilities - Diamond, ISIS, ...

• Strictly a data archive - no metadata, very limited search - but useful

• Data pulled off tape when needed, to staging or for download

• Archive goes back lifetime of Diamond
EXAMPLE: ZENODO?

- Funded through EU / openAIRE
- Intended to be free at the point of access for depositors and users
- Allows but does not mandate metadata
- Allows curation via communities
- Provides DOI, search facilities etc.
- Provides open API -> very useful
ZENODO FOR CRYSTALLOGRAPHY?

- General archive - so “mandatory data” does not include everything for e.g. CIF - but it could
- Not optimised for our use case - no scope for adding structured metadata
- If we started uploading 100,000 raw data sets / year someone would notice
- Great for “one off” type uploads
Raw diffraction data for structure of SARS-CoV-2 main protease with PCM-0102575 (ID: mpro-x1351 / PDB: 5R5K).

Sample: Mpro-x1351
Fluo: 3.50kW-1
θ Start: 90.0°
θ Stop: 90.0°
θ Overlap: 6°
No. Images: 400
Resolution: 1.8Å
Wavelength: 0.911Å
Exposure: 0.24E11
Transmission: 100.00%
Beamsize: 60x60µm
Type: SAD

Comment: (1/200,12,184) X-ray centering boxes: [29.14Å (42), 29.14Å (42), 24.0µm (24)], Aperture: 70µm
WHAT DO WE WANT?

- Ability to annotate raw data with processing output, full experiment metadata, sample material etc.
- Link to published structure - but not mandatory? Publishing unsuccessful data very interesting
- Validation to ensure that the data correspond to the claimed structure
- Facility to automate publication and update
CURATION

- Currently zenodo communities have “light touch” curation - largely done as a hobby by folks - but librarian is a vocation / job

- PDB, CSD etc. have professional curators and annotators - adding value to the raw data and the archive

- Critical to ensure the standards are defined

- Critical to ensure the standards are maintained

- Critical to ensure people are helped as users and depositors
COSTS

- Disks are cheap, small, portable
- Can buy one for every visit to DLS for a small cost compared with other consumables
- Obviously data not public, but can consider making public if someone asks…
COSTS

Data storage is expensive

• disks die
• technology changes - try finding a firewire port in 2021
• failure / accidents happen
• if you have not tried to read the data, assume the worst
COSTS

Data storage is expensive

- disks die
- technology changes - try finding a firewire port in 2021
- failure / accidents happen
- if you have not tried to read the data, assume the worst
REAL COSTS - STORAGE

- Azure as an example - probably priced in a realistic manner
- One “visit” / “shift” ~ 4TB
- £2.48 / month cheapest storage cost - £300 over 10 years
- 10 years time you’ll be paying to store 60 visits worth of data... and the data won’t be getting smaller

### Data storage prices pay-as-you-go

All prices are per GB per month.

<table>
<thead>
<tr>
<th></th>
<th>Premium</th>
<th>Hot</th>
<th>Cool</th>
<th>Archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>First 50 terabyte (TB) / month</td>
<td>£0.11180 per GB</td>
<td>£0.0135 per GB</td>
<td>£0.00746 per GB</td>
<td>£0.00074 per GB</td>
</tr>
<tr>
<td>Next 450 TB/month</td>
<td>£0.11180 per GB</td>
<td>£0.0129 per GB</td>
<td>£0.00746 per GB</td>
<td>£0.00074 per GB</td>
</tr>
<tr>
<td>Over 500 TB/month</td>
<td>£0.11180 per GB</td>
<td>£0.0124 per GB</td>
<td>£0.00746 per GB</td>
<td>£0.00074 per GB</td>
</tr>
</tbody>
</table>

### Azure Storage Reserved Capacity

Azure Storage Reserved Capacity helps you lower your data storage cost by committing to one year or three years of Azure Storage. Reserved capacity can be purchased in increments of 100 TB and 1 PB sizes for 1-year and 3-year commitment durations. All prices are per month. For more information, please see documentation.

<table>
<thead>
<tr>
<th></th>
<th>1-year reserved</th>
<th>3-year reserved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hot</td>
<td>Cool</td>
</tr>
<tr>
<td>100 TB/month</td>
<td>£1,152</td>
<td>£626</td>
</tr>
<tr>
<td>1 PB/month</td>
<td>£11,217</td>
<td>£6,096</td>
</tr>
</tbody>
</table>
REAL COSTS - CURATION

- Different shape to storage - up front / one off rather than annual
- Partly amenable to automation - but still work to verify the data match the publication etc.
- Highly dependent on the goals of the curated data archive - the higher the value, the higher the staff costs
WORKED EXAMPLE

• 25 GB data set -> £2 to store for 10 years at cheapest rate

• Processing time to validate - 15 minutes on 16 core machine - £0.2 (low priority cloud resource)

• People cost to verify data - 5 minutes at £25 / hour -> £2

• Overall about £5 / data set (€6 / $7)

• Taking the data out will cost about £0.75 - £1.25 a go…
WHO PAYS?

- Scientist - reader - traditional manuscript model
- Creator - new “open access” model
- Facility (common in e.g. radio astronomy)
- 3rd party
WHO PAYS? CHALLENGES

- Scientist - reader - traditional manuscript model - additional expense for hard pressed labs - also implies that publishers have control over your data (same as papers)

- Creator - new “open access” model - additional costs again to labs, though not impossible - advantage that it scales - but lab funding is transient

- Facility (common in e.g. radio astronomy) - very expensive as we don’t know what data will be important, also have to support many disciplines

- 3rd party - how are we going to persuade someone of the need?
HYBRID MODEL

- Data archive - facility / zenodo / azure (assumed to be reliable, may or may not provide DOI) - need not be specialised for crystallography

- Metadata archive - with the publication of the structure - has DOI - is curated and contains a reference back to the raw data (build into CSD / PDB) - see e.g. extensions to imgCIF to allow references to HDF5 raw data
SHOULD WE PAY?

• £6 / data set is / is not good value

• How much does it cost to reproduce the data?

• How much value will the data have? Will anyone ever look at it?
CONCLUSIONS

• Archiving raw experimental data perfectly possible - see Zenodo - easy even

• Defining a standard perfectly possible - see achievements in CIF / mmCIF / PDB etc. - making it part of publication process excellent way of encouraging people

• Deciding who should do the archiving is hard - and who should pay for it, how long the archive should live etc.

• Hybrid model of separating the data archiving from the metadata and curation more likely to meet the community need - just need to ensure link is bidirectional
ACKNOWLEDGEMENTS

• Diamond / STFC IT folks for keeping ICAT running and useful

• Diamond / STFC staff, users

• Commenters on Twitter for raising useful questions

• NeXus / imgCIF (& Herbert Bernstein) for standards definitions