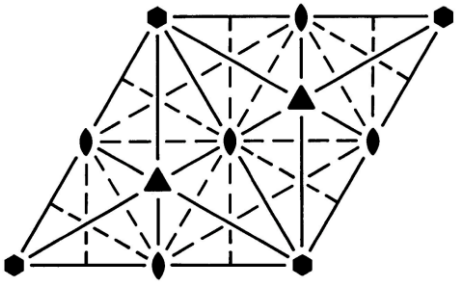
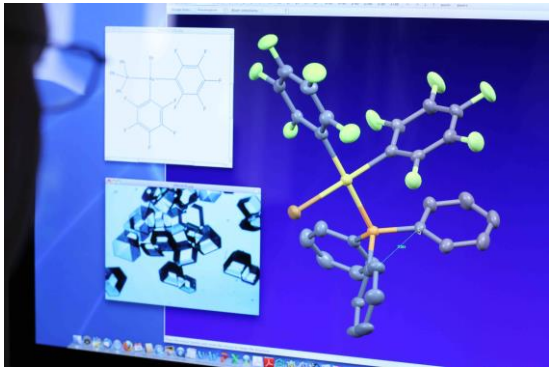
University of Zurich[UZH]

Department of Chemistry

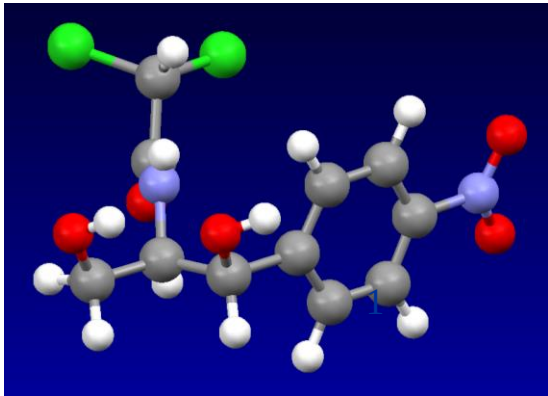
Data refereeing and editing in chemical crystallography; the Acta Cryst. C experience



Anthony Linden



ECM-32 Workshop:
Data Science Skills in Publishing
Vienna, 18 August, 2019



The science is in the data

- Funding bodies are increasingly demanding...
  - publication only in Open Access journals;
    Plan S is the extreme case
  - there must now be open access to all primary data
- How to implement?
- Do funding bodies understand the requirements needed to achieve this?

FAIR principles for Open Research Data

- Findable, Accessible, Interoperable, Reusable
- FAIR alone is not sufficient; archived data should be true facts
- FACT and FAIR are needed for reproducibility
- FACTs require validation tools

Publication and data management in crystallography

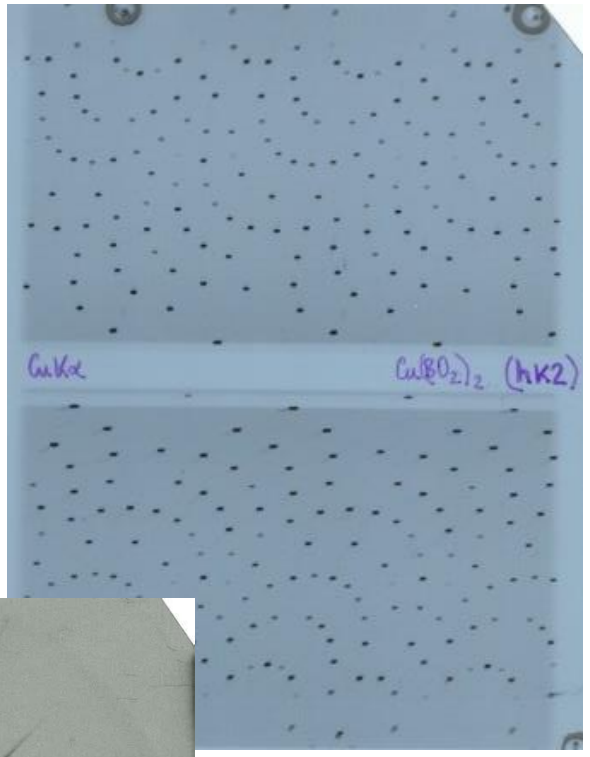
- Crystallography is a very digital/numerical science
- Relatively easy to adhere to FAIR if suitable tools and repositories are available
- IUCr journals have a long history of data management
  - structure factors
  - CIF
  - validation
  - input reflection data & refinement instructions
  - the next phase: archiving raw diffraction images

Handling of raw diffraction data

- In the beginning...

  ...there was film

- Much of diffraction space recorded. Non-Bragg features recognised early on.
- How many films still exist today – and where?
- Were films ever requested in article review?
- Intensities estimated by eye



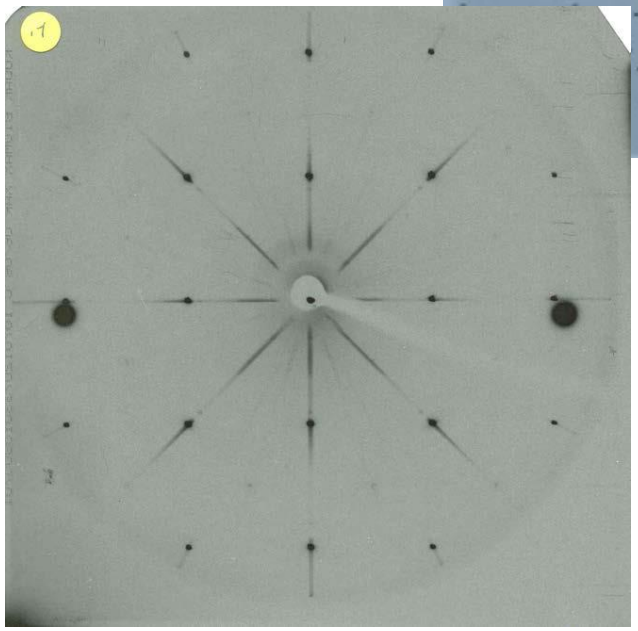
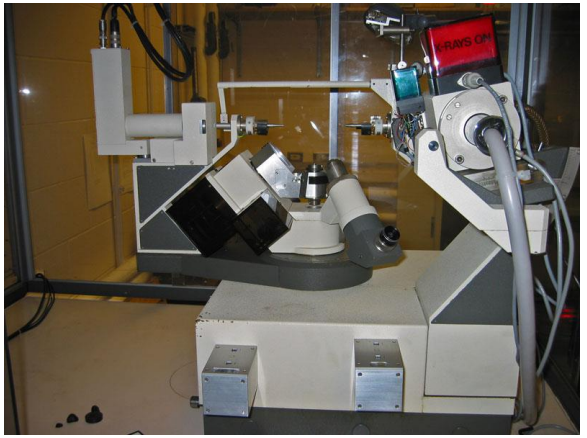


Image from Martin Martinez-Ripoll http://www.xtal.iqfr.csic.es/Cristalografia/

Serial diffractometers, point detectors

- Faster & automated
- Non-Bragg scattering generally ignored.
  Who knew it was there? Twinning overlooked?
- Intensities accurately determined if
  background featureless
- Raw data looked like this...




```
HG9207  C19H32N2S4      11-MAR-92   17:05:59     Directory:
     9.3947    12.1129    20.3796     90.013    102.379     89.992     2265.23

STANDARD DATA FOLLOWS:
 0     1  -3  -4  -7    70.84     .36 0 1  10769   40   63   .01    .00  25.61  12.80  35.13  52.74      8 1122
 0     2   0  -5   9    63.85     .35 0 1   8979   48   58   .02    .00  25.13  12.56  42.63 -99.43      8 1122
 0     3   0   5   9    64.08     .35 0 1   8999   27   60   .04    .00  25.13  12.57 -41.55 -88.13      8 1122
 2     4   0   0 -19      .00     .00 0 1    166   89  120   .04    .00  39.65  19.83   -.53  86.27      8 1160
 1     5   0   0 -18    21.69     .53 0 1    733   18   28   .05    .00  37.49  18.75   -.53  86.27      8 1152
 2     6   0   0 -17      .00     .00 0 1    158  148   46   .05    .00  35.33  17.67   -.53  86.27      8 1148
 1     7   0   0 -16    24.11     .49 0 1   1052   31   37   .06    .00  33.19  16.59   -.53  86.27      8 1141
 2     8   0   0 -15      .00     .00 0 1    159   42   91   .06    .00  31.06  15.53   -.53  86.27      8 1137
 2     9   0   0 -14     3.46    1.74 0 1    171   36   34   .07    .00  28.95  14.48   -.53  86.27      8 1133
```

Integrated intensity over scan Background left Background right

Serial diffractometers, point detectors

- Raw data...

```
HG9207  C19H32N2S4      11-MAR-92   17:05:59     Directory:
     9.3947    12.1129    20.3796     90.013    102.379     89.992     2265.23

STANDARD DATA FOLLOWS:
 0     1  -3  -4  -7    70.84     .36 0 1  10769   40   63   .01    .00  25.61  12.80  35.13  52.74     8 1122
 0     2   0  -5   9    63.85     .35 0 1   8979   48   58   .02    .00  25.13  12.56  42.63 -99.43     8 1122
 0     3   0   5   9    64.08     .35 0 1   8999   27   60   .04    .00  25.13  12.57 -41.55 -88.13     8 1122
 2     4   0   0 -19      .00     .00 0 1    166   89  120   .04    .00  39.65  19.83   -.53  86.27     8 1160
 1     5   0   0 -18    21.69     .53 0 1    733   18   28   .05    .00  37.49  18.75   -.53  86.27     8 1152
 2     6   0   0 -17      .00     .00 0 1    158  148   46   .05    .00  35.33  17.67   -.53  86.27     8 1148
 1     7   0   0 -16    24.11     .49 0 1   1052   31   37   .06    .00  33.19  16.59   -.53  86.27     8 1141
 2     8   0   0 -15      .00     .00 0 1    159   42   91   .06    .00  31.06  15.53   -.53  86.27     8 1137
 2     9   0   0 -14     3.46    1.74 0 1    171   36   34   .07    .00  28.95  14.48   -.53  86.27     8 1133
```

- Processed data...

```
  0   0 -18  3588.21  132.70
  0   0 -17 -1470.07  962.33
  0   0 -16  4629.75  139.99
```

F^2 $\sigma(F^2)$

Serial diffractometers, point detectors

- Raw data...

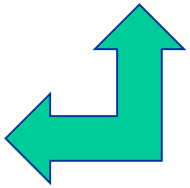
```
HG9207  C19H32N2S4      11-MAR-92   17:05:59     Directory:
     9.3947    12.1129    20.3796     90.013    102.379     89.992     2265.23

STANDARD DATA FOLLOWS:
 0     1  -3  -4  -7    70.84     .36 0 1  10769   40   63   .01    .00  25.61  12.80  35.13  52.74     8 1122
 0     2   0  -5   9    63.85     .35 0 1   8979   48   58   .02    .00  25.13  12.56  42.63 -99.43     8 1122
 0     3   0   5   9    64.08     .35 0 1   8999   27   60   .04    .00  25.13  12.57 -41.55 -88.13     8 1122
 2     4   0   0 -19      .00     .00 0 1    166   89  120   .04    .00  39.65  19.83   -.53  86.27     8 1160
 1     5   0   0 -18    21.69     .53 0 1    733   18   28   .05    .00  37.49  18.75   -.53  86.27     8 1152
 2     6   0   0 -17      .00     .00 0 1    158  148   46   .05    .00  35.33  17.67   -.53  86.27     8 1148
 1     7   0   0 -16    24.11     .49 0 1   1052   31   37   .06    .00  33.19  16.59   -.53  86.27     8 1141
 2     8   0   0 -15      .00     .00 0 1    159   42   91   .06    .00  31.06  15.53   -.53  86.27     8 1137
 2     9   0   0 -14     3.46    1.74 0 1    171   36   34   .07    .00  28.95  14.48   -.53  86.27     8 1133
```

- Processed data...

```
  0   0 -18  3588.21  132.70
  0   0 -17 -1470.07  962.33
  0   0 -16  4629.75  139.99
```

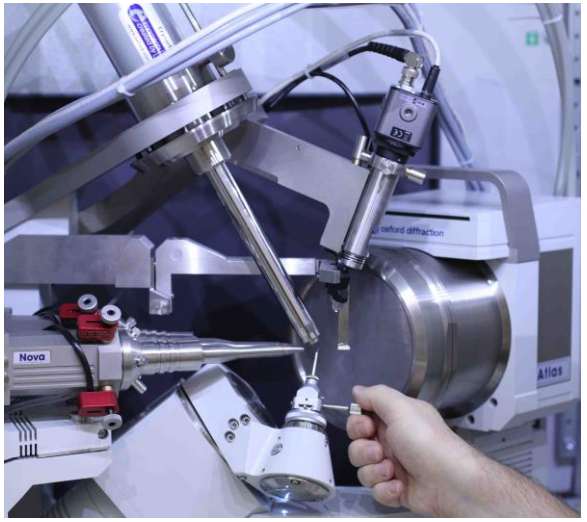
F^2 $\sigma(F^2)$



What happens in between?
Which gets archived?
Which was used for review?

The CCD era

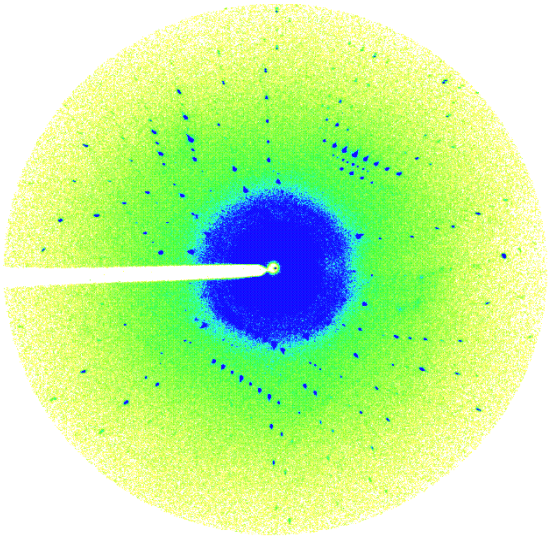
- 1994 – the Bruker SMART 1000 CCD detector introduced. Revolution in data collection speed – we also see the non-Bragg features again!
- Diffractometer and structure solution/refinement software developed quickly (better GUIs)
- The technique becomes more accessible to a wider range of users
- Many more structures coming out of labs, many more “lay crystallographers” doing structures
- Less time to cogitate over results and their correctness or implications
- CIF and electronic validation are thus crucial



Large detectors, beam lines & X-FELs

- Dectris Eiger X 16M: 18,139,650 pixels @ 133 Hz (frames/sec)
- Eiger X 500K: up to 9 kHz
- Massive amounts of data generated rapidly
- The infrastructure today can handle it!

- But where & how do we store and archive all this?
- The raw data are no longer ascii text files
- How are the data used in the review process?



Review and publication of data

- In the early days, many journals generally published:
  - crystal data
  - atomic coordinates & atomic displacement parameters (ADPs)

C(4)	854 (4)	1956 (4)	3892 (3)	83 (3)	40 (4)	27 (2)	2 (3)	8 (2)	−10 (2)	
C(5)	653 (4)	2712 (4)	3207 (3)	69 (4)	65 (4)	22 (2)	8 (3)	4 (2)	−12 (2)	
C(6)	1692 (5)	4191 (5)	2746 (3)	99 (5)	89 (5)	27 (2)	8 (4)	29 (2)	−1 (3)	
C(7)	3309 (5)	3216 (5)	4032 (4)	70 (5)	80 (5)	46 (3)	13 (4)	15 (3)	−4 (3)	
C(8)	2346 (5)	1721 (4)	5034 (3)	76 (5)	60 (4)	32 (2)	14 (3)	2 (2)	1 (2)	
C(9)	119 (5)	1737 (4)	4374 (3)	84 (5)	51 (4)	35 (2)	−10 (3)	12 (3)	−14 (2)	

Acta Cryst. B 1976

  - observed and calculated structure factors [F_o, F_c, $\sigma(F_o)$]

0	1	-4	378	353	-1	6	0	1662	1684	-1	5	3	839	890	-1	3	6	1729	1850	* -1	-2	10	94	49	-2	9	-2	581	526
0	1	4	2513	2429	1	6	0	581	574	-1	5	-3	1270	1228	-1	3	-6	1337	1431	1	3	-10	756	773	-2	10	-2	148	146
0	2	4	2689	2591	1	7	0	837	850	1	5	3	1865	1893	1	3	-6	1378	1325	-1	3	-10	959	1076	2	0	3	1297	1206
0	2	-4	2892	2826	-1	8	0	2135	2142	-1	6	-3	2473	2434	1	4	-6	1108	1091	1	4	-10	986	1079	-2	0	3	1621	1655
0	3	4	540	549	1	8	0	391	350	1	6	-3	702	699	* 1	4	6	94	74	-1	4	-10	648	719	2	1	3	1919	1910
0	3	-4	1756	1755	-1	9	0	919	929	1	6	3	216	236	-1	4	6	567	571	-1	5	-10	270	295	2	-1	3	378	384
0	4	4	324	243	1	9	0	675	742	-1	6	3	635	607	-1	4	-6	635	621	1	5	-10	337	335	-2	1	3	1162	969
0	4	-4	1919	1931	-1	10	0	216	165	1	7	3	189	165	1	5	-6	1000	979	1	6	-10	135	75	2	1	-3	2284	2073
0	5	-4	310	286	1	0	-1	5270	5750	* -1	7	3	94	142	-1	5	6	324	326	2	0	0	1121	1033	2	2	-3	1419	1336
0	5	4	986	986	1	0	1	2527	2486	-1	7	-3	2270	2222	-1	5	-6	1094	1070	-2	1	0	932	864	-2	2	3	648	612
0	6	-4	3689	3667	1	1	-1	1013	929	1	7	-3	1148	1152	1	5	6	1054	1059	2	1	0	2878	2903	-2	2	-3	986	984
0	6	4	229	252	1	1	1	3473	3324	1	8	-3	513	504	-1	6	-6	432	399	-2	2	0	932	988	2	2	3	324	437
0	7	-4	2000	1914	-1	1	1	2554	2602	1	8	3	554	570	1	6	-6	243	160	2	2	0	162	179	2	3	-3	567	518
0	7	4	135	191	-1	1	-1	2135	2020	-1	8	3	702	637	1	6	6	581	651	2	3	0	864	807	-2	3	3	2689	2522
* 0	8	4	94	67	1	2	1	729	716	-1	8	-3	283	341	-1	6	6	419	517	-2	3	0	2135	2181	-2	3	-3	2216	2230
0	8	-4	946	936	-1	2	1	1716	1584	* 1	9	-3	94	87	-1	7	6	243	262	-2	4	0	689	661	2	3	3	1527	1536

Acta Cryst. B 1968

Review and publication of data

- What did a reviewer learn from printed F_o/F_c tables?
  - at best an outlier was noticed almost by accident
  - many $F_o > F_c$ might have suggested twinning
- How accessible are the archived data for further work?
- What corrections were applied and were they suitable or the best possible?
- New ideas or tools for data processing cannot be applied to processed data
- Better to archive the original raw data, but where are they now?

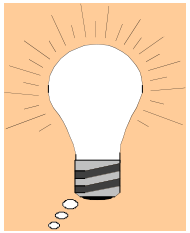
The demise of SF tables, coordinates & ADP data

- By the 1970s, most journals did not print SF tables – archived on microfilm (*e.g. Inorg. Chem*. in 1971)
- Most journals then decided they did not want to receive SF tables at all *(Inorg. Chem*. in 1992). Reviewers had to do without!
- Manuscripts still submitted in paper form in triplicate at that time! Review involved looking only at the submitted printed information.

- Journals instructed authors to take responsibility for holding onto SF tables for at least 12 months! How FAIR is that?
- With the advent of CIF, journals soon stopped printing tables of atomic coordinates, ADPs and some crystal data

The demise of SF tables, coordinates & ADP data

- Even with CIF, many journals were not interested in the SF listings
- IUCr journals are one of the few stalwarts in this regard.
  Submission and archiving of SF listings has never been interrupted
- The CSD only started accepting SF listings around 2010-2012
- The CSD also did not store ADPs in the actual database
  – a legacy issue addressed only very recently
  – original versions of the deposited CIFs can be requested
  – what about for depositions before CIF?
- Before CIF, exchange of data between labs could involve significant file reformatting – poor Interoperability

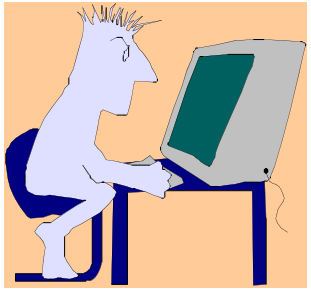


CIF and the (public) Internet

- Both came into existence around the same time (1991)
- CIF: S.R. Hall, F.H. Allen, I.D. Brown, *Acta Cryst*. **1991**, *A47*, 655
- The CIF standard allowed easy interchange, archival & retrieval of data and ensured uniform information content
- The Internet allowed easy transmission of data (horrible for large files, initially). The web simplified search & retrieval.
- Journals slowly moved into electronic manuscript submission & publication

CIF and online submission & publishing

- *Acta Cryst. C* was an early leader in electronic submission of manuscripts, completely within a CIF, in 1993
- IUCr journals went online in January 2000
- IUCr journals stopped printing issues in January 2014



- Getting CIF accepted quickly by the community was no mean feat
  - George Sheldrick released *SHELXL*-93 with CIF output
  - mmCIF was also adopted well by the macromolecular community
  - some software in the powder community lagged somewhat behind

Working with early CIFs

- Only contained experimental details & results of the structure determination
- Atomic coordinates, ADPs, derived geometry, crystal & experimental data
- Structure factors, merged, only available in a separate file
- No refinement instructions (clues about refinement strategy)
- No uncorrected or unmerged input reflections
- No raw diffraction images
- Difficult to investigate a structure further, such as aspects the original authors had no interest in
- Can't go back if new ideas or techniques come along, such as merged light atom data preventing testing absolute structure by new methods
- Can't undo/redo absorption corrections, *SQUEEZE*, twin handling, restraints

Validation

- Up to 1997, *Acta Cryst. C* co-editors received a handwritten “validation” report generated by the technical staff in Chester to help with the workload, but otherwise had to check structures manually

- 1997: Syd Hall, the IUCr office and Ton Spek collaborated on the development of electronic validation of CIFs
- Formally introduced for *Acta Cryst. C* submissions at the start of 1998
- Unpopular at the beginning

checkCIF

A service of the
International Union of Crystallography

Validation

- Allowed maintenance of best practice standards, which seemed to have been slipping at the time
- Allowed detection of possible oversights by authors in everything from data collection to model finalization
- Helped “inexperienced” practitioners
- Validation of structure factors introduced only in 2010
- Inclusion of refinement instructions/strategy requested from 2010

CIF in the *Acta Cryst. C* review process

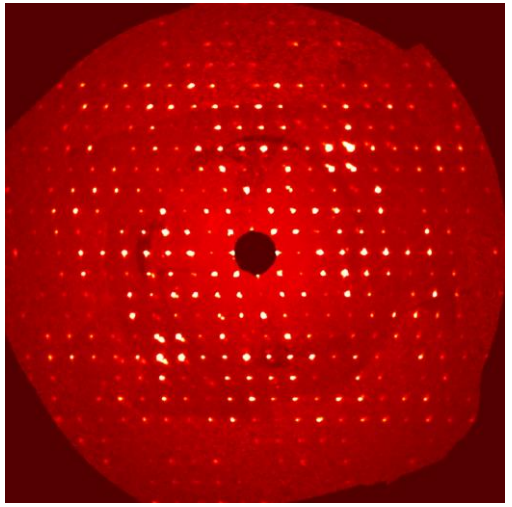
- CIF enabled electronic validation, easier archival, searching & retrieval
- Co-editors now work to the same standards for data & results quality
- Co-editors can focus more on the scientific content and quality of papers
- Early days of *checkCIF* did not assess...
  - structure factors – F_o/F_c submitted in CIF format as a separate file
  - refinement strategy (input instructions)
  - input reflection data
  - raw diffraction data

Data in the *Acta Cryst. C* review process

- If validation or inspection of the CIF threw up questions, Co-editors could ask authors for their refinement instructions and even the input reflection file – perhaps to try an alternative refinement strategy for the authors

- Almost never asked for diffraction images
  - how to open images if they did not have the same instrument?
  - transmission of a full data set can be cumbersome

- Reconstructions of reciprocal lattice layer precession images (unwarping) occasionally requested during review; they can be revealing about the crystal quality, presence of twinning, *etc*. Generally not archived, though.



- And we trusted that everyone was doing the best job possible for their level of expertise...

Fraudulent structures in *Acta Cryst. E* ☹

- Large number discovered around 2009
- Often the structures of slightly different chemical compounds were presented based on the same, or slightly modified, reflection data
- The existing *checkCIF* based only on the CIF, not the SF data, could detect these only if one looked critically at the minor alerts
- Examination of the SF data was more revealing
- *checkCIF* was then extended to include examination of the structure factor files
- This revealed that sometimes, even for legitimate structures, the submitted SF file did not match the CIF!
  - a problem related to having to submit independent files

Data in the *Acta Cryst. C* review process

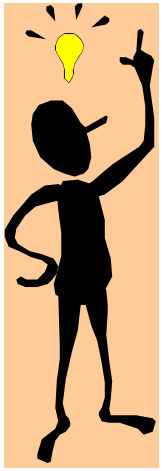
- *Acta Cryst. C & E* recommended (2010), then required (2012) the refinement instructions to be embedded in the CIF.
  - authors have forgotten about documenting non-standard refinements!
  - helps to see if the strategy used by authors, especially for difficult structures, is optimal.
  - helps others understand or reproduce the authors' result if they are interested in extending the work, methods research, *etc*.
- *SHELXL*-2014 automatically embeds the refinement instructions and input (unmerged) reflection data in the CIF –
  - easy for authors; everything now in the one file
  - some authors complained about managing such "large" CIFs!

Raw diffraction data in the review process today

- Not currently used routinely for *Acta Cryst. C* – maybe in special cases
- Not required during manuscript submission
- Most useful for review if automated *checkCIF*-like software is available
- Plans to validate raw powder diffraction data have not yet come to fruition
- Deposition of processed diffraction data, but not raw images, is now routine for macromolecular structure deposition in the PDB, probably initially instigated as a result of some "misinterpretations" of data
- From July 2019, the PDB only accepts mmCIF, no longer PDB format

Deposition and archiving of raw diffraction data

- Needs significant infrastructure
- Needs to be an enduring and actively curated archive
- Needs extensive metadata
- Needs to be searchable (doi preferred)
- Repositories at least at an institutional level or national level
- Archiving just within a single research group is less ideal – significant administrative effort required, data may not be maintained well, be less accessible and lost when a research group dissolves



Deposition and archiving of raw diffraction data

- The capacity and accessibility of digital archives are now reducing the hurdle to reviewing and preserving large quantities of raw data
- But authors might initially be reluctant...
- In the *Acta Cryst. C* experience, authors often express annoyance when new requirements are introduced
- Busy authors will embrace open access for primary data and FAIR principles if suitable repositories are available, and the deposition, access and data extraction procedures are as routine, transparent, automatic and effort-free as possible.
- The IUCr is working towards this goal

New depositories for macromolecular data

- Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC, https://www.proteindiffraction.org)
- Structural Biology Grid Consortium (SBGrid, https://sbgrid.org)
- Australian Store.Synchrotron facility (https://store.synchrotron.org.au)
- Depository for X-ray lasers (CXIDB, https://www.cxidb.org)
- General research data repositories such as Zenodo (CERN, https://zenodo.org)

Information resources

- Update from the IUCr Committee on Data, May 2019:

J.R. Helliwell, W. Minor, M.S. Weiss, E.F. Garman, R.J. Read, J. Newman, M.J. van Raaij, J. Hajdu, E.N. Baker, Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography, *IUCrJ*, **2019**, *6*, 341-343

doi: 10.1107/S2052252519005918

"IUCr Journals are now taking the lead by encouraging authors to provide a doi for their deposited original raw diffraction data…"

- For a more detailed discourse, see also*:*

L.M.J. Kroon-Batenburg, J.R. Helliwell, B. McMahon, T.C. Terwilliger, Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements, *IUCrJ*, **2017**, *4*, 87-99

doi: 10.1107/S2052252516018315