The element of trust: validating and valuing crystallographic data

Brian McMahon[1], John R. Helliwell[2] and James R. Hester[3]

[1]International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK. E-mail: bm@iucr.org

[2]School of Chemistry, University of Manchester, Manchester M13 9PL, England

[3]Australian Nuclear Science and Technology Organisation, New Illawarra Road, Lucas Heights, NSW 2234, Australia

AsCA 2018 CRYSTAL 32
December 2-5 2018 | Auckland, New Zealand

The record of science

- The International Union of Crystallography is committed to the highest quality in preserving the record of science

  - specifically in its publishing activities

Crystallography Journals Online

Acta Crystallographica Section A: Foundations and Advances

Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials

Acta Crystallographica Section C: Structural Chemistry

Acta Crystallographica Section D: Biological Crystallography

Acta Crystallographica Section E: Crystallographic Communications

Acta Crystallographica Section F: Structural Biology Communications

Journal of Applied Crystallography

Journal of Synchrotron Radiation

IUCrJ

The IUCr and data

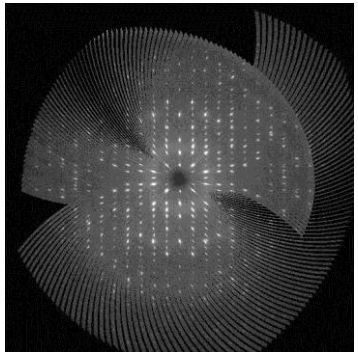
- Data also form an essential part of the record of science
- But *crystallographic data* are many and varied…

The data zoo

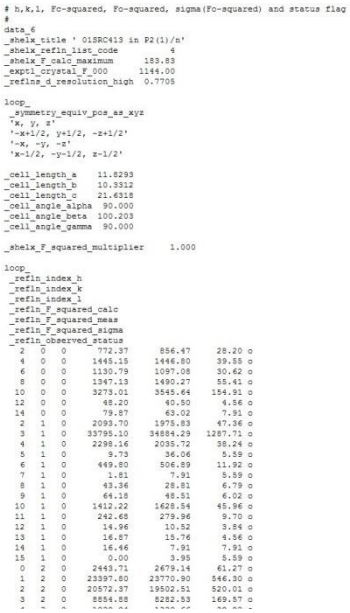
Data can mean any or all of:

1. raw measurements from an experiment
2. processed numerical observations
3. derived structural information
4. variable parameters in the experimental set-up or numerical modelling and interpretation
5. bibliographic and linking information

We make no fundamental distinction between **data** and **metadata** – metadata *are* data (though they may be of secondary interest to the current focus of attention).



(1)



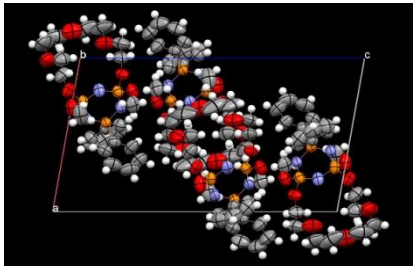
```
# h,k,l, Fo-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_refln_list_code          4
_shelx_F_calc_maximum       183.83
_exptl_crystal_F_000       1144.00
_refins_d_resolution_high  0.7705

loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, y+1/2, -z+1/2'
 '-x, -y, -z'
 'x-1/2, -y-1/2, z-1/2'

_cell_length_a    11.8293
_cell_length_b    10.3312
_cell_length_c    21.6318
_cell_angle_alpha  90.000
_cell_angle_beta  100.203
_cell_angle_gamma  90.000

_shelx_F_squared_multiplier     1.000

loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0       772.37       856.47       28.20 o
   4   0   0      1445.15      1446.80       39.85 o
   6   0   0      1130.79      1097.08       30.62 o
   8   0   0      1347.13      1490.27       55.41 o
  10   0   0      3273.01      3545.64      154.91 o
  12   0   0        48.20        40.50        4.56 o
  14   0   0        79.87        63.02        7.91 o
   2   1   0      2093.70      1975.83       47.36 o
   3   1   0     33795.10     34884.29     1287.71 o
   4   1   0      2298.16      2035.72       38.24 o
   5   1   0         9.73        36.06        5.59 o
   6   1   0       449.80       506.89       11.92 o
   7   1   0         1.81         7.91        5.59 o
   8   1   0        43.36        28.81        6.79 o
   9   1   0        64.18        48.51        6.02 o
  10   1   0      1412.22      1628.54       45.96 o
  11   1   0       242.68       279.96        9.70 o
  12   1   0        14.96        10.52        3.56 o
  13   1   0        16.87        15.76        4.56 o
  14   1   0        16.46         7.91        7.91 o
  15   1   0         0.00         3.95        5.59 o
   0   2   0      2443.71      2679.14       61.27 o
   1   2   0     23397.80     23770.90      546.50 o
   2   2   0     20572.37     19502.51      520.01 o
   3   2   0      8854.88      8282.53      169.57 o
```

(2)

Table 1. *Selected geometric parameters* (Å, °)

Fe1—C9	2.030 (4)	Fe1—C7	2.049 (4)
Fe1—C5	2.036 (3)	Fe1—C11	2.053 (4)
Fe1—C12	2.038 (4)	S1—C1	1.693 (3)
Fe1—C13	2.038 (4)	N1—C1	1.315 (4)
Fe1—C4	2.038 (3)	N2—C1	1.345 (4)
Fe1—C8	2.041 (4)	N2—N3	1.387 (4)
Fe1—C10	2.042 (4)	N3—C2	1.290 (4)
Fe1—C6	2.048 (4)		
C1—N2—N3	118.2 (3)	N2—C1—S1	120.1 (2)
C2—N3—N2	116.9 (3)	N3—C2—C4	115.6 (3)
N1—C1—N2	117.0 (3)	N3—C2—C3	125.2 (3)
N1—C1—S1	122.8 (2)	C4—C2—C3	119.3 (3)

(3)



(4)

Acta Crystallographica Section C
Crystal Structure Communications
ISSN 0108-2701

L-Histidyl-L-serine 3.7-hydrate: water channels in the crystal structure of a polar dipeptide

Carl Henrik Görbitz

Department of Chemistry, University of Oslo, Oslo, Norway
Correspondence e-mail: c.h.gorbitz@kjemi.uio.no

Received 20 August 2010
Accepted 24 September 2010
Online 8 October 2010

Dipeptides may form nanotubular structures with pore diameters in the range 3.2–10 Å. These compounds normally contain at least one and usually two hydrophobic residues, but L-His-L-Ser hydrate, $C_9H_{14}N_4O_4{\cdot}3.7H_2O$, with two hydrophilic residues, forms large polar channels filled with ordered as well as disordered water molecules.

(5)

The element of trust

Trust in

- Data transmission/exchange
  - Crystallographic Information File (1991)
- Data consistency
  - checkCIF for derived (coordinate) data (1998)
  - checkCIF including structure factors (2007)
- Data provenance
  - Diffraction data deposition (2011-2017)
- The science within the data

IUCr activities relating to data

- 1991 Crystallographic Information Framework
- 1998 *checkCIF*
- 2011 Diffraction data deposition
- 2017 Committee on Data (CommDat)

CIF

- Developed as uniform file format for data exchange

CIF basics

```
data_99107abs
_chemical_name_systematic
; 3-Benzo[b]thien-2-yl-5,6-dihydro-1,4,2-oxathiazine 4-
    oxide
;
_chemical_name_common              ?
_chemical_formula_iupac           'C11 H9 N O2 S2'
_chemical_formula_moiety          'C11 H9 N O2 S2'
_chemical_formula_sum             'C11 H9 N O2 S2'
_chemical_formula_weight           251.31
loop_
 _atom_site_label
 _atom_site_type_symbol
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
 _atom_site_U_iso_or_equiv
 _atom_site_adp_type
S4   S  0.32163(7)   0.45232(6) 0.52011(3) 0.04532(13) Uani
S11  S  0.39642(7)   0.67998(6) 0.29598(2) 0.04215(12) Uani
O1   O  -0.00302(17) 0.67538(16) 0.47124(8) 0.0470(3) Uani
O4   O  0.2601(2)    0.28588(16) 0.50279(10) 0.0700(5) Uani
H5A  H  0.1284       0.4834 0.6221 0.060 Uiso
H5B  H  0.1861       0.6537 0.5908 0.060 Uiso
```

- Simple text file (ASCII character set)
- Tags (data names) begin with underscore
- Data values are text strings representing numbers or text
- White space delimits (strings containing white space must be quoted)
- loop_ before set of data names sets up a tabular relationship
- Data name + data value = **data item**

Drivers for CIF – standardisation (of format)

J. Appl. Cryst. (1992). **25**, 455–459

DIFRAC, single-crystal diffractometer output-conversion software. By H. D. Flack, *Laboratoire de Cristallographie, University of Geneva, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland* and E. Blanc and D. Schwarzenbach, *Institut de Cristallographie, University of Lausanne, BSP Dorigny, CH-1015 Lausanne, Switzerland*

(*Received* 25 *October* 1991; *accepted* 9 *January* 1992)

Abstract

Software is described that converts single-crystal diffractometer output files as produced by maufacturers' software into a standardized instrument-independent form consisting of a clear, complete, well documented record of the sample and the diffraction measurements performed upon it. Information not already available in the manufacturers' diffractometer files is obtained in an interactive question-and-answer session. The software is written in a modular way. Available modules can deal with Enraf–Nonius CAD-4, Philips PW1100 and Siemens $P2_1$ single-crystal diffractometers and produce output in CIF or SCFS format.

Introduction

Users of several models of four-circle single-crystal diffractometers may well have been struck by the diversity of form and content of the data files generated by diffractometer-manufacturers' software. The form of the file can create difficulties in its transfer to other computing equipment and further renders the corresponding computer programs specific to a certain type or types of diffractometer.

“

The difficulties in the content of the files manifest themselves principally by ***the paucity of the available information*** *necessitating additional input to the data-treatment software (e.g. type of radiation, wavelength of radiation, scan width, …). The problem has become aggravated in recent times by the rapid development in electronic data exchange. ... The advent of machine-readable submission for publication and data-base and supplementary-material deposition further highlights* ***the problem of missing data****.*

”

CIF

- Developed as uniform file format for data exchange
- Adoption of dictionaries → ‘ontologies’

Drivers for CIF – standardisation (of terminology), completeness

Version: **CIF Dictionary (Core 1991)**

_atom_site_aniso_label *(char)*

Anisotropic atomic displacement parameters are usually looped in a separate list. If this is the case, this code must match the **_atom_site_label** of the associated atom coordinate list and conform with the same rules described in **_atom_site_label**.

Appearance in list: yes.

_atom_site_aniso_type_symbol *(char)*

This **_atom_type_symbol** code links the anisotropic atom parameters to the atom type data associated with this site and must match one of the **_atom_type_symbol** codes in this list.

Appearance in list: yes. If looped, **_atom_site_aniso_label** must be present in the same list.

_atom_site_aniso_U_11
_atom_site_aniso_U_12
_atom_site_aniso_U_13
_atom_site_aniso_U_22
_atom_site_aniso_U_23
_atom_site_aniso_U_33 *(numb)*

These are the standard anisotropic atomic displacement components which appear in the structure factor term: $\exp(-2\pi^2 \sum_i \sum_j U_{ij} h_i h_j a_i^* a_j^*)$. The components may be entered in any order.

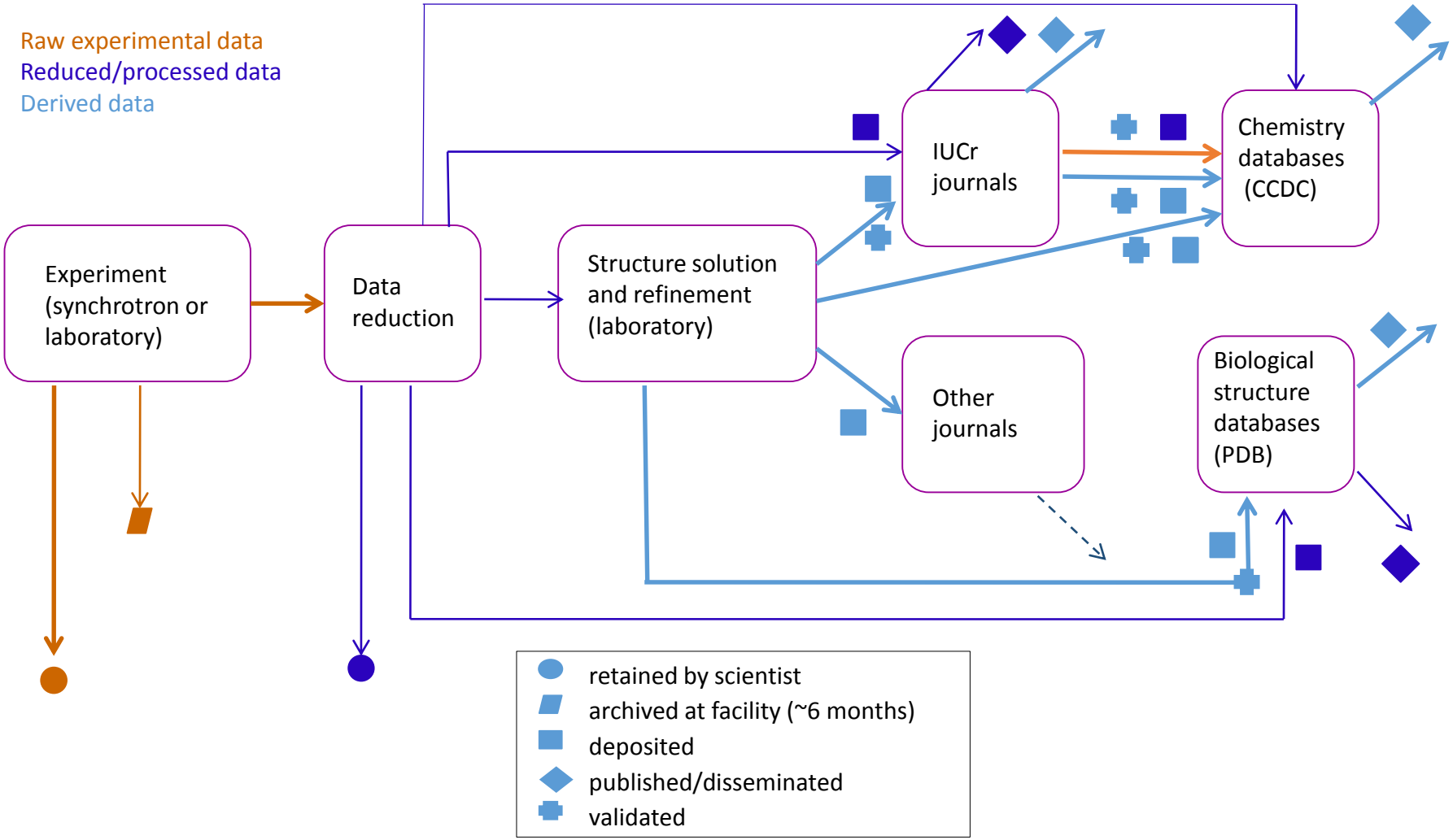
Appearance in list: yes. If looped, **_atom_site_aniso_label** must be present in the same list. E.s.d. expected: yes. Default e.s.d. value: 0.0. The units extensions are: ' ' (ångströms squared *1.0) '_pm' (picometres squared /10000.) '_nm' (nanometres squared *100.).

```
data_atom_site_aniso_label
    _name                       '_atom_site_aniso_label'
    _type                        char
    _list                        yes
    _definition
;               Anisotropic atomic displacement parameters are usually looped in
                a separate list. If this is the case, this code must match the
                _atom_site_label of the associated atom coordinate list and
                conform with the same rules described in _atom_site_label.
;

data_atom_site_aniso_type_symbol
    _name                       '_atom_site_aniso_type_symbol'
    _type                        char
    _list                        yes
    _list_identifier            '_atom_site_aniso_label'
    _definition
;               This _atom_type_symbol code links the anisotropic atom
                parameters to the atom type data associated with this site and
                must match one of the _atom_type_symbol codes in this list.
;

data_atom_site_aniso_U_
    loop_ _name                 '_atom_site_aniso_U_11'
                                '_atom_site_aniso_U_12'
                                '_atom_site_aniso_U_13'
                                '_atom_site_aniso_U_22'
                                '_atom_site_aniso_U_23'
                                '_atom_site_aniso_U_33'
    _type                        numb
    _list                        yes
    _list_identifier            '_atom_site_aniso_label'
    _esd                         yes
    _esd_default                 0.0
    loop_ _units_extension
          _units_description
          _units_conversion      ' '   'Angstroms squared'  *1.0
                                 '_pm'  'picometres squared' /10000.
                                 '_nm'  'nanometres squared' *100.
    _definition
;               These are the standard anisotropic atomic displacement
                components which appear in the structure factor term:
                exp(-2pi^2^ sum~i~ sum~j~ U~ij~ h~i~ h~j~ a*~i~ a*~j~).
                The components may be entered in any order.
;
```

CIF

- Developed as uniform file format for data exchange
- Adoption of dictionaries → ‘ontologies’
- Use in publication (small-cell-parameter structures)
- Use in databases (PDB, CSD)

A coherent information flow



CIF

- Developed as uniform file format for data exchange
- Adoption of dictionaries → ‘ontologies’
- Use in publication (small-cell-parameter structures)
- Use in databases (PDB, CSD)
- Unifying framework for definition of metadata

CIF as a metadata catalogue for crystallography

checkCIF

- Began as IUCr journals in-house suite of validation/consistency checks

Checking for completeness

checkcif.iucr.org/cgi-bin/checkcif_hkl.pl

Datablock: I

```
Bond precision:       = 0.0000 A                    Wavelength=0.71073
Cell:       a=1             b=1             c=1
            alpha=90        beta=90         gamma=90
Temperature: 0 K
                  Calculated                      Reported
Volume            1                               738.55(7)
Space group       P 21 21 21                      ?
Hall group        : P 2ac 2ab                     ?
Moiety formula
Sum formula
Mr                0.00                            152.15
Dx,g cm-3         0.000                           1.368
Z                 1                               4
Mu (mm-1)         0.000                           0.105
F000              0.0                             0.0
F000'             0.00
h,k,lmax                                          6,11,18
Nref                                              871
Tmin,Tmax         0.978,0.983
Tmin'             0.950
Correction method= Not given
Data completeness=              Theta(max)= 26.000
R(reflections)= 0.0438( 706)         wR2(reflections)= 0.1113( 871)
S = 1.085              Npar= 101
```

The following ALERTS were generated. Each ALERT has the format
test-name_ALERT_alert-type_alert-level.
Click on the hyperlinks for more details of the test.

Alert level A

SYMM001_ALERT_1_A _symmetry_cell_setting is missing
The cell setting should be one of the following
* triclinic
* monoclinic
* orthorhombic
* tetragonal
* rhombohedral
* trigonal
* hexagonal
* cubic

The following tests will not be performed.
SYMMS_01,SYMMS_02

PLAT198_ALERT_1_A Missing _diffrn_ambient_temperature Datum .... Please Add
PLAT801_ALERT_4_A Cell Data Missing, Incomplete or Out-of-Order Please Check

Alert level G

PLAT005_ALERT_5_G No Embedded Refinement Details Found in the CIF Please Do !
PLAT104_ALERT_1_G The Reported Crystal System is Inconsistent with P212121 Check

Checking for completeness

checkcif.iucr.org/cgi-bin/checkcif_hkl.pl

Datablock: I

```
Bond precision:       = 0.0000 A                    Wavelength=0.71073
Cell:        a=1              b=1              c=1
             alpha=90         beta=90          gamma=90
Temperature: 0 K
                    Calculated                    Reported
Volume              1                             738.55(7)
Space group         P 21 21 21                    ?
Hall group          : P 2ac 2ab                   ?
Moiety formula
Sum formula
Mr                  0.00                          152.15
Dx,g cm-3           0.000                         1.368
Z                   1                             4
Mu (mm-1)           0.000                         0.105
F000                0.0                           0.0
F000'               0.00
h,k,lmax                                          6,11,18
Nref                                              871
Tmin,Tmax           0.978,0.983
Tmin'               0.950
Correction method= Not given
Data completeness=                Theta(max)= 26.000
R(reflections)= 0.0438( 706)          wR2(reflections)= 0.1113( 871)
S = 1.085               Npar= 101
```

The following ALERTS were generated. Each ALERT has the format
test-name_ALERT_alert-type_alert-level.
Click on the hyperlinks for more details of the test.

Alert level A

SYMM001_ALERT_1_A _symmetry_cell_setting is missing
The cell setting should be one of the following
* triclinic
* monoclinic
* orthorhombic
* tetragonal
* rhombohedral
* trigonal
* hexagonal
* cubic
The following tests will not be performed.
SYMMS_01,SYMMS_02
PLAT198_ALERT_1_A Missing _diffrn_ambient_temperature Datum .... Please Add
PLAT801_ALERT_4_A Cell Data Missing, Incomplete or Out-of-Order Please Check

Alert level G

PLAT005_ALERT_5_G No Embedded Refinement Details Found in the CIF Please Do !
PLAT104_ALERT_1_G The Reported Crystal System is Inconsistent with P212121 Check

Checking for consistency

Alert level A

```
PLAT150_ALERT_1_A Volume as Calculated Differs from that Given ...    738.55 Ang-3
PLAT198_ALERT_1_A Missing _diffrn_ambient_temperature  Datum ....     Please Add
PLAT701_ALERT_1_A Bond   Calc   1.384(5), Rep   1.368(5), Dev..      3.20 Sigma
          C3     -C2          1.555  1.555 .......    #   1 Check
PLAT701_ALERT_1_A Bond   Calc   1.527(5), Rep   1.496(4), Dev..      6.20 Sigma
          C8     -C7          1.555  1.555 .......    #  13 Check
PLAT702_ALERT_1_A Angle  Calc   121.9(3), Rep   122.9(3), Dev..      3.33 Sigma
          O2     -C8    -O1       1.555  1.555  1.555    #  13 Check
PLAT707_ALERT_1_A D...A  Calc   2.732(4), Rep   2.676(4), Dev..     14.00 Sigma
          O3     -O2          1.555  4.745 .......    #  19 Check
PLAT707_ALERT_1_A D...A  Calc   2.747(3), Rep   2.689(3), Dev..     19.33 Sigma
          O1     -O3          1.555  2.665 .......    #  19 Check
```

Alert level B

```
PLAT701_ALERT_1_B Bond   Calc   1.382(4), Rep   1.371(4), Dev..      2.75 Sigma
          C2     -C1          1.555  1.555 .......    #   4 Check
PLAT701_ALERT_1_B Bond   Calc   1.383(5), Rep   1.368(5), Dev..      3.00 Sigma
          C6     -C5          1.555  1.555 .......    #   6 Check
PLAT701_ALERT_1_B Bond   Calc   1.395(4), Rep   1.385(5), Dev..      2.50 Sigma
          C5     -C4          1.555  1.555 .......    #   9 Check
PLAT702_ALERT_1_B Angle  Calc   123.6(3), Rep   122.9(3), Dev..      2.33 Sigma
          O2     -C8    -C7       1.555  1.555  1.555    #  14 Check
```

Alert level C

```
PLAT046_ALERT_1_C Reported Z, MW and D(calc) are Inconsistent ....    1.329 Check
PLAT068_ALERT_1_C Reported F000 Differs from Calcd (or Missing)...    Please Check
PLAT242_ALERT_2_C Low    'MainMol' Ueq as Compared to Neighbors of      C8 Check
PLAT340_ALERT_3_C Low Bond Precision on  C-C Bonds ..............    0.0045 Ang.
PLAT355_ALERT_3_C Long   O-H (X0.82,N0.98A)  O3       - H3A     .     1.01 Ang.
PLAT701_ALERT_1_C Bond   Calc   1.195(5), Rep   1.186(4), Dev..      1.80 Sigma
          C8     -O2          1.555  1.555 .......    #  11 Check
PLAT702_ALERT_1_C Angle  Calc   120.9(3), Rep   120.3(3), Dev..      2.00 Sigma
          C3     -C2    -C1       1.555  1.555  1.555    #   4 Check
PLAT702_ALERT_1_C Angle  Calc   118.5(3), Rep   117.9(3), Dev..      2.00 Sigma
          C6     -C5    -C4       1.555  1.555  1.555    #  10 Check
PLAT702_ALERT_1_C Angle  Calc   120.0(3), Rep   120.5(3), Dev..      1.67 Sigma
          C4     -C5    -C7       1.555  1.555  1.555    #  12 Check
PLAT702_ALERT_1_C Angle  Calc   119.1(3), Rep   119.5(3), Dev..      1.33 Sigma
          C2     -C1    -C6       1.555  1.555  1.555    #  16 Check
PLAT702_ALERT_1_C Angle  Calc   120.7(3), Rep   121.1(3), Dev..      1.33 Sigma
          C3     -C4    -C5       1.555  1.555  1.555    #  19 Check
```

Alert level G

```
PLAT005_ALERT_5_G No Embedded Refinement Details Found  in the CIF    Please Do !
PLAT007_ALERT_5_G Number of Unrefined Donor-H Atoms ..............       2 Report
PLAT721_ALERT_1_G Bond   Calc    1.04000, Rep    1.02990 Dev...      0.01 Ang.
          C1     -H1          1.555  1.555 .......    #  14 Check
PLAT721_ALERT_1_G Bond   Calc    0.92000, Rep    0.89950 Dev...      0.02 Ang.
          C7     -H7B         1.555  1.555 .......    #  17 Check
PLAT721_ALERT_1_G Bond   Calc    1.01000, Rep    0.98500 Dev...      0.02 Ang.
          O3     -H3A         1.555  1.555 .......    #  18 Check
PLAT721_ALERT_1_G Bond   Calc    0.97000, Rep    0.95040 Dev...      0.02 Ang.
          O1     -H1A         1.555  1.555 .......    #  19 Check
PLAT722_ALERT_1_G Angle  Calc     119.00, Rep     117.90 Dev...      1.10 Degree
          C2     -C1    -H1       1.555  1.555  1.555    #  17 Check
PLAT722_ALERT_1_G Angle  Calc     119.00, Rep     117.40 Dev...      1.60 Degree
          C8     -C7    -H7B      1.555  1.555  1.555    #  25 Check
PLAT725_ALERT_2_G D-H    Calc    1.01000, Rep    0.98000 Dev...      0.03 Ang.
          O3     -H3A         1.555  1.555 .......    #  19 Check
PLAT725_ALERT_2_G D-H    Calc    0.97000, Rep    0.95000 Dev...      0.02 Ang.
          O1     -H1A         1.555  1.555 .......    #  19 Check
PLAT726_ALERT_2_G H...A  Calc    1.74000, Rep    1.71000 Dev...      0.03 Ang.
          H3A    -O2          1.555  4.745 .......    #  19 Check
PLAT726_ALERT_2_G H...A  Calc    1.78000, Rep    1.74000 Dev...      0.04 Ang.
          H1A    -O3          1.555  2.665 .......    #  19 Check
```

A transcription error in a cell parameter (b = 9.3092 Å instead of 9.0392 Å) generates many inconsistencies and anomalies in bond geometry.

checkCIF

- Began as IUCr journals in-house suite of validation/consistency checks
- Consolidated with PLATON and offered as public service

https://checkcif.iucr.org

checkCIF

- Began as IUCr journals in-house suite of validation/consistency checks
- Consolidated with PLATON and offered as public service
- Adopted by other journal publishers and databases
- Needs extension to validate other techniques
- Needs longer-term maintenance
- Model for other applications (‘*checkCIF* for raw data’)

Checking for scientific reasonableness

- Q: What should I do to fix my checkCIF ‘errors’?
- A: (*ideally*) NOTHING!

Of course, real errors at the experimental or refinement stage do need to be fixed, but residual alerts should indicate outliers from expected model behaviour. These should be explained.

  - Experimental difficulties (*e.g.* poor crystal)
  - Modelling difficulties
  - Novel science

Checking for scientific reasonableness

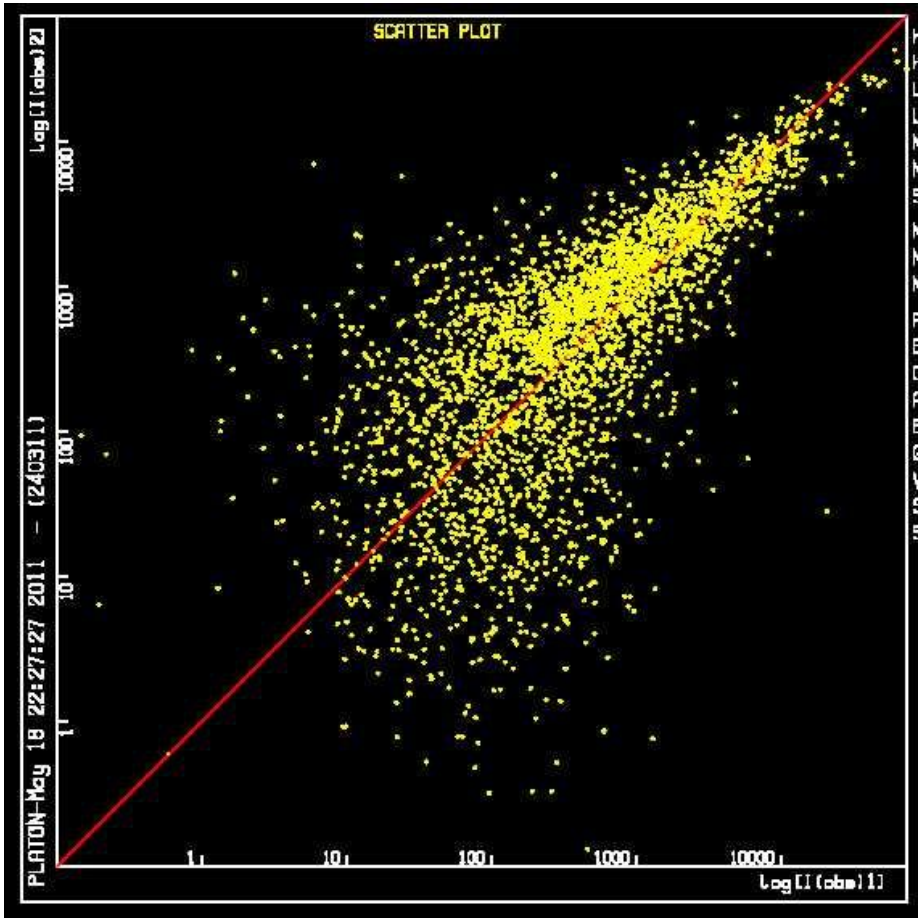
Here is an example of a completed VRF

```
_vrf_PLAT_213_global
;
PROBLEM:  Atom C(6B) has ADP max/min ratio .....      5.20
RESPONSE:
          Atom C6 of the ring (B) was found to be disordered;
          see _publ_section_exptl_refinement
;
```

The completed VRF should be inserted in the CIF after the first datablock identifier (*i.e.* after the `data_something` line that indicates the start of a CIF data block).

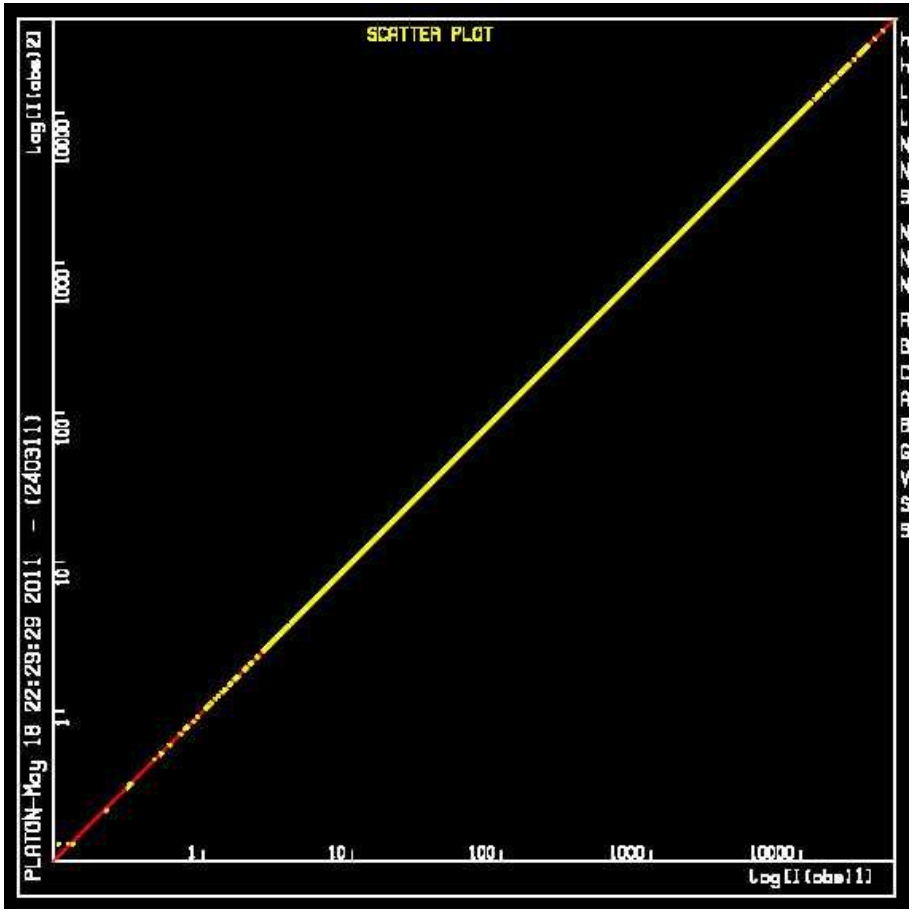
Ideally, the VRF should be added to the CIF using *publCIF*.

Checking for provenance



In the period 2007-2009 over 100 structure determinations were published in *Acta Cryst. E* that were subsequently retracted. There is some suspicion that these were fraudulent results; at best they were evidence of serious errors in handling and managing experimental data. Each structure determination was supported by data sets that were broadly consistent with the reported structure. In several cases there were clusters of level C alerts in the *checkCIF* reports that were slightly odd, but they generally were within the range of error that Co-editors had seen and accepted in other structures. The nature of the problem only became apparent when comparing the experimental data sets deposited with each structure against previous submitted structures. This scatter plot is typical of the degree of correlation that one would normally see between comparative data points in two data sets from distinct (albeit related) chemical species.

Checking for provenance



However, in some cases the correlation was unexpectedly strong. Investigation showed that the experimental data sets associated with different chemical structures were, in fact, subsets of a single original data set. Even the most charitable interpretation of this suggests egregious errors in the handling of the experimental data by the contributing authors. In this case the errors were found by cross-comparison of distinct data sets, a procedure made possible by the IUCr's insistence on archiving experimental data for small-cell-parameter structures. Such errors could be reduced by strong protocols that embed provenance metadata within the data handling chain (i.e. that unequivocally link raw data to reduced data sets to the ultimate structure publication).

COMCIFS

- Created at Beijing Congress (1993) to maintain CIF standard
- Supervises and commissions dictionaries

CIF dictionaries (COMCIFS)

- Crystallographic Core (coreCIF) – 1991
- Crystallographic Restraints – 2011
- Crystallographic Powder Diffraction (pdCIF) – 1997
- Modulated and Composite Structures (msCIF) – 2002
- Multipole Electron Density (rhoCIF) – 2003
- Crystallographic Twinning – 2014
- Magnetic Structures (magCIF) – 2016
- Lattice topology – *in development*
- Crystallographic Symmetry (symCIF) – 2001
- Diffraction Images (imgCIF) – 2000
- Crystallographic Macromolecular Structure – 1997

COMCIFS

- Created at Beijing Congress (1993) to maintain CIF standard
- Supervises and commissions dictionaries
- mmCIF family of dictionaries managed by wwPDB since 2006

CIF dictionaries (wwPDB)

- Crystallographic Macromolecular Structure – 1997
- PDB Exchange Dictionary (PDBx/mmCIF) – 1997 and ongoing
- Integrative/Hybrid (I/H) methods – 2017
- 3DEM Extension Dictionary – 2004
- NMRSTAR Dictionary – 2013
- Biological Small Angle Scattering– 1998
- Model Archive Extension Dictionary – 2018
- BIOSYNC Extension Dictionary – 2000
- NMR Exchange Format Dictionary – 2016

COMCIFS

- Created at Beijing Congress (1993) to maintain CIF standard
- Supervises and commissions dictionaries
- mmCIF family of dictionaries managed by wwPDB since 2006
- Encourages software development
- Occasional workshops

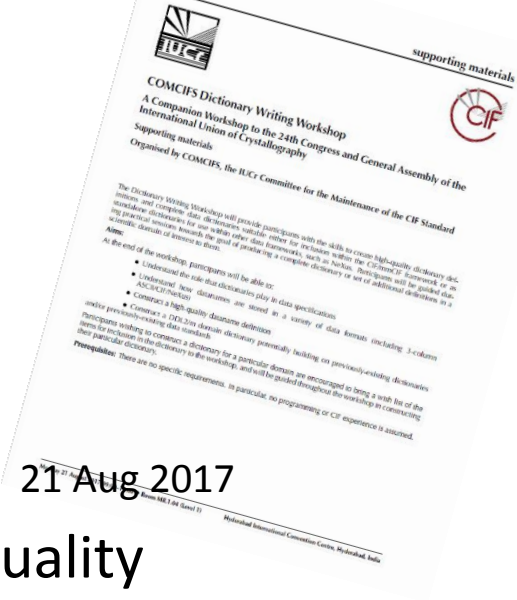
COMCIFS activities

- Warwick Workshop (ECM 28) 23-24 Aug 2013
  - Objective: Introduction to the CIF2 extended syntax; development of an API to permit community development of open-source libraries and tools; introduction to DDLm, a dictionary definition language supporting algorithmic methods; discussion of dREL, a prototyping methods evaluator language; tutorials and demonstrations of *JsCifBrowser*, a CIF2 implementation in JavaScript; approaches to CIF dictionary authoring.
  - https://www.iucr.org/resources/cif/comcifs/workshop-2013

COMCIFS activities

- Warwick Symposium (ECM 28) 25 Aug 2013
  - Objective: to celebrate and develop the role of the CIF standard in crystallographic information and data management.
  - https://www.iucr.org/resources/cif/comcifs/symposium-2013

COMCIFS activities


supporting materials

COMCIFS Dictionary Writing Workshop

A Companion Workshop to the 24th Congress and General Assembly of the International Union of Crystallography

supporting materials

Organised by COMCIFS, the IUCr Committee for the Maintenance of the CIF Standard

- Dictionary writing workshop (IUCr XXIV) 21 Aug 2017
  - Objective: to provide skills to create high-quality dictionary definitions and complete data dictionaries suitable either for inclusion within the CIF/mmCIF framework or as standalone dictionaries for use within other data frameworks, such as NeXus.

DDDWG

Diffraction Data Deposition Working Group

- Convened at Madrid Congress (2011) to assess desirability and feasibility of routine deposition of raw diffraction images
- Initial scepticism about usefulness of routine data deposition
- Community engagement by DDDWG (forums, sessions at national/regional meetings) raised awareness
- Scoping papers in *Acta D* and pilot projects demonstrated feasibility of non-centralised archiving (linking via DOI)

DDDWG publications

- Terwilliger, T. C. (2014). Archiving raw crystallographic data. *Acta Cryst.* D**70**, 2500–2501.
- Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). Experiences with making diffraction image data available: what metadata do we need to archive? *Acta Cryst.* D**70**, 2502–2509.
- Meyer, G. R., Aragao, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). Operation of the Australian Store.Synchrotron for macromolecular crystallography. *Acta Cryst.* D**70**, 2510–2519.
- Guss, J. M. & McMahon, B. (2014). How to make deposition of images a reality. *Acta Cryst.* D**70**, 2520–2532.
- Terwilliger, T. C. & Bricogne, G. (2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. *Acta Cryst.*D**70**, 2533–2543.
- Baker, E. N. (2017). Data archiving and availability in an era of open science. *IUCrJ* **4**, 1–2.
- Grabowski, M. & Minor, W. (2017). Sharing Big Data. *IUCrJ* **4**, 3–4.
- Kroon-Batenburg, L. M. J., Helliwell, J. R., McMahon, B. & Terwilliger, T. C. (2017). Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ*, **4**, 87–99.
- Bruno, I., Gražulis, S., Helliwell, J. R., Kabekkodu, S. N., McMahon, B. & Westbrook, J. (2017). Crystallography and Databases. *Data Sci. J.* **16**, p. 38.
- Helliwell, J. R., McMahon, B., Guss, J. M. & Kroon-Batenburg, L. M. J. (2017). The science is in the data. *IUCrJ* **4**, 714–722.

DDDWG

Diffraction Data Deposition Working Group

- Convened at Madrid Congress (2011) to assess desirability and feasibility of routine deposition of raw diffraction images
- Initial scepticism about usefulness of routine data deposition
- Community engagement by DDDWG (forums, sessions at national/regional meetings) raised awareness
- Scoping papers in *Acta D* and pilot projects demonstrated feasibility of non-centralised archiving (linking via DOI)
- Workshops (especially second) focused on metadata requirements

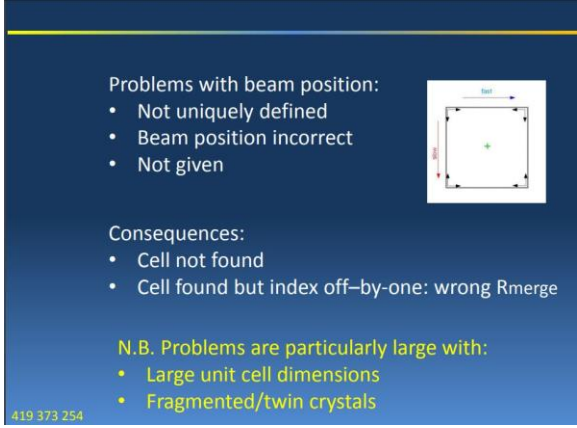
DDDWG activities

- Bergen Workshop (ECM 27) 6 Aug 2012
  - Objective: To help frame a policy to be drafted by the IUCr DDD WG on raw diffraction data deposition for final approval by the IUCr Executive Committee.
  - https://www.iucr.org/resources/data/dddwg/bergen-workshop

DDDWG activities

- Rovinj Workshop (ECM 29) 22-23 Aug 2015
  - Objective: to define the necessary metadata to be captured and deposited alongside experimental diffraction images so that such raw data may be subsequently re-evaluated or re-used in more detailed scientific studies. The workshop will also explore the metadata requirements of other structural experimental techniques used by crystallographers.
  - https://www.iucr.org/resources/data/dddwg/rovinj-workshop

DDDWG activities



- New Orleans Workshop (ACA 2017) 26 May 2017
  - Objective: (1) What every experimentalist needs to know about recording essential metadata of raw diffraction data
    (sample preparation and characterization; correct recording of instrument axes, correction factors, calibration; attention to diffuse scattering or other interesting "metadata“);
    (2) Research Data Management policy mandates and requirements on Principal Investigators (PIs)
    (metadata standardization; data repositories; primary data linking to publications).
  - https://www.iucr.org/resources/data/dddwg/new-orleans-workshop

DDDWG

Diffraction Data Deposition Working Group

- Convened at Madrid Congress (2011) to assess desirability and feasibility of routine deposition of raw diffraction images
- Initial scepticism about usefulness of routine data deposition
- Community engagement by DDDWG (forums, sessions at national/regional meetings) raised awareness
- Scoping papers in *Acta D* and pilot projects demonstrated feasibility of non-centralised archiving (linking via DOI)
- Workshops (especially second) focused on metadata requirements
- Gradual acceptance of desirability of archiving some data (difficult crystals, difficult solutions, diffuse scattering)

DDDWG recommendations (1)

- **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 should obey 'FAIR' principles (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 URL) as the most sustainable way to identify and locate a raw diffraction data set.
- An **archive of raw diffraction data** sets for **currently unsolved crystal structures** should be pursued.
- An **archive of raw diffraction data** sets **showing significant diffuse scattering** should be pursued.

DDDWG recommendations (2)

- Workshops for **research data management training** for the community should continue and be sponsored and organised by the IUCr.
- There should be continued regular checking by the IUCr Executive Committee of the progress of the **IUCr Commissions** logging of their **raw diffraction data metadata**.
- **Archived raw** diffraction **data** should be **automatically validated** wherever possible via a '**checkcif for raw data** approach', and be peer reviewed where necessary, at the minimum to include core metadata: beam centre of diffraction image, wavelength, wavelength bandpass (pink beam case), orientation of all axes, pixel sizes, detector position and orientations.
- Jointly with the IUCr Commission on Crystallographic Computing, the IUCr should pursue **reproducibility of science objectives which require open source software and accurate versioning**.
- IUCr should engage with vendors and the World Data System to promote the **certification of raw diffraction data standards**.

DDDWG recommendations (3)

- IUCr's CommDat … should continue the **directory of data archives** by adding any new data archives that are established in future. [L. M. J. Kroon-Batenburg *et al.* (2017) *IUCrJ*, **4**, 87-99.]
- IUCr should invite the community to alert CommDat of further **case studies** that document the value of archiving of raw diffraction data. [Current case study examples are included in J. R. Helliwell *et al.* (2017) *IUCrJ*, **4**, 714-722.]
- IUCr recognises that **metadata for the sample** are clearly vital for all the IUCr Commissions (and are especially diverse in small angle scattering), and whose standardised descriptions should be actively pursued by the Commissions.
- CommDat should regularly monitor the **evolution of technology** as the pace of change in data measurement rates, and of **metadata logging**, with new detectors, computer hardware, networks and electronic laboratory notebooks is especially notable.
- IUCr should actively **support** the neutron, synchrotron and X-ray laser **facilities** in their raw data archiving activities.

DDDWG

Diffraction Data Deposition Working Group

- Convened at Madrid Congress (2011) to assess desirability and feasibility of routine deposition of raw diffraction images
- Initial scepticism about usefulness of routine data deposition
- Community engagement by DDDWG (forums, sessions at national/regional meetings) raised awareness
- Scoping papers in *Acta D* and pilot projects demonstrated feasibility of non-centralised archiving (linking via DOI)
- Workshops (especially second) focused on metadata requirements
- Gradual acceptance of desirability of archiving some data (difficult crystals, difficult solutions, diffuse scattering)
- Engagement with FAIR principles and wider movements towards reproducible and Open Science

CommDat

IUCr Committee on Data

- Formed in 2017 at Hyderabad Congress as long-term (standing) committee
- Subsumes activities of DDDWG
- Continues data-related interests of the now discontinued Committee on Crystallographic Databases and of the Committee on Electronic, Publishing, Dissemination and Storage of Information
- Data liaison body with CODATA and ICS

Current CommDat projects

- *checkCIF* for raw data (Loes Kroon-Batenburg, James Hester)

checkCIF for raw data

- Continuing project of IUCr CommDat
- Currently focused on (X-ray) diffraction images
- Aims to move towards standards of
  - Metadata completeness (*e.g*. axis definitions, beamstop position) to facilitate interpretation
  - Consistency
  - Provenance (to verify genuine research inferences; important in publication peer-review process)

checkCIF for raw data (images)

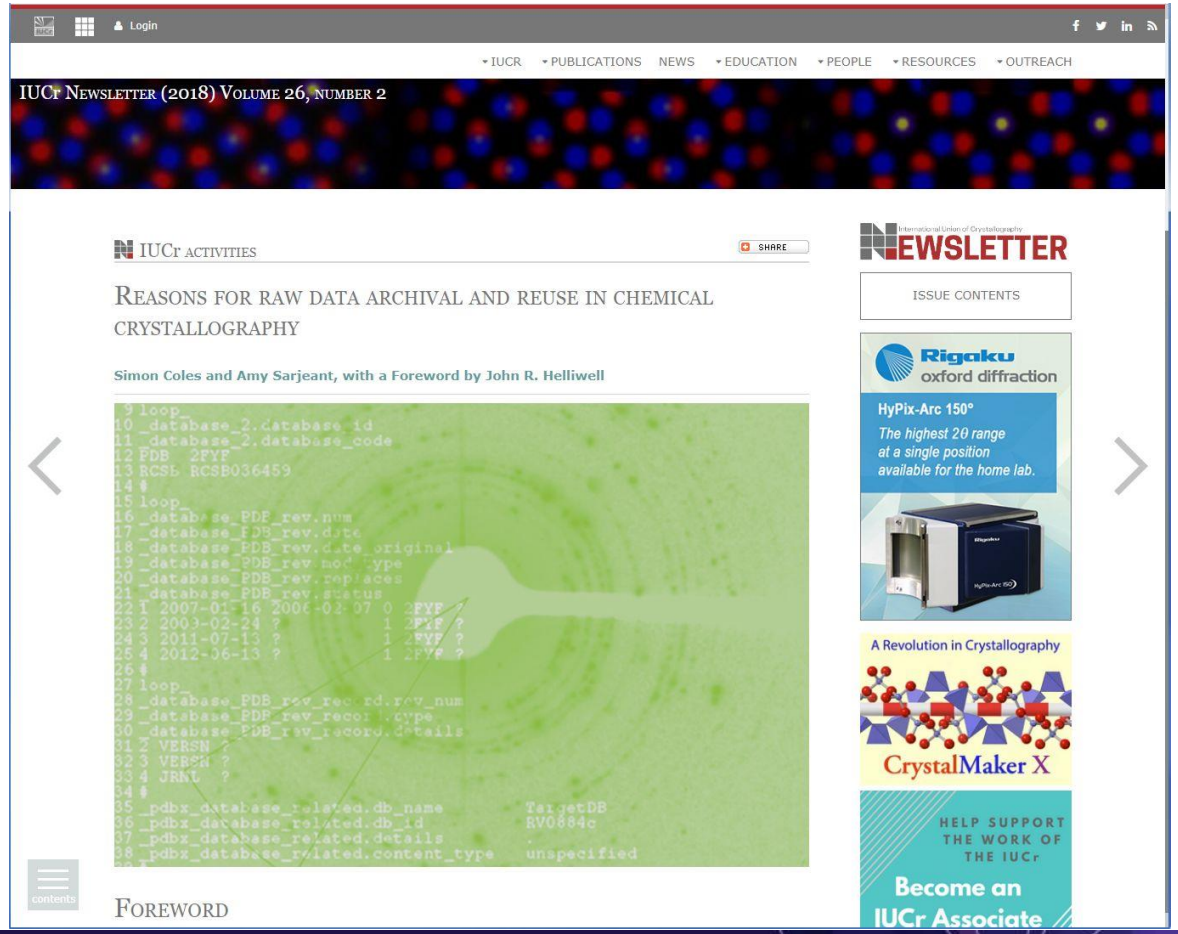
(1) Has the relationship of the detector axes to the image axes been specified?
(2) Has the relationship of the detector axes to laboratory coordinates been specified?
(3) Has the relationship of the goniometer axes to laboratory coordinates been specified (if the sample moves)?
(4) Is the specified beam centre located at a realistic position? (In a shadowed region)
(5) Do calculated HKL spot positions match with observed spots?
(6) Does the geometry specified in (1), (2), (3) "make sense"?
(7) Is the wavelength available?

'Pre-tests' for any dataset that is submitted for verification.
(1) Has the format of all data files been specified (*e.g.* ADSC, NXMX, imgCBF, Bruker)?
(2) Has the meaning of items in the data files been specified (by reference to a CIF dictionary)?

Current CommDat projects

- *checkCIF* for raw data (Loes Kroon-Batenburg, James Hester)
- Need for archiving of raw data in chemical crystallography (Simon Coles, Amy Sarjeant)

Raw data archiving in chemical crystallography

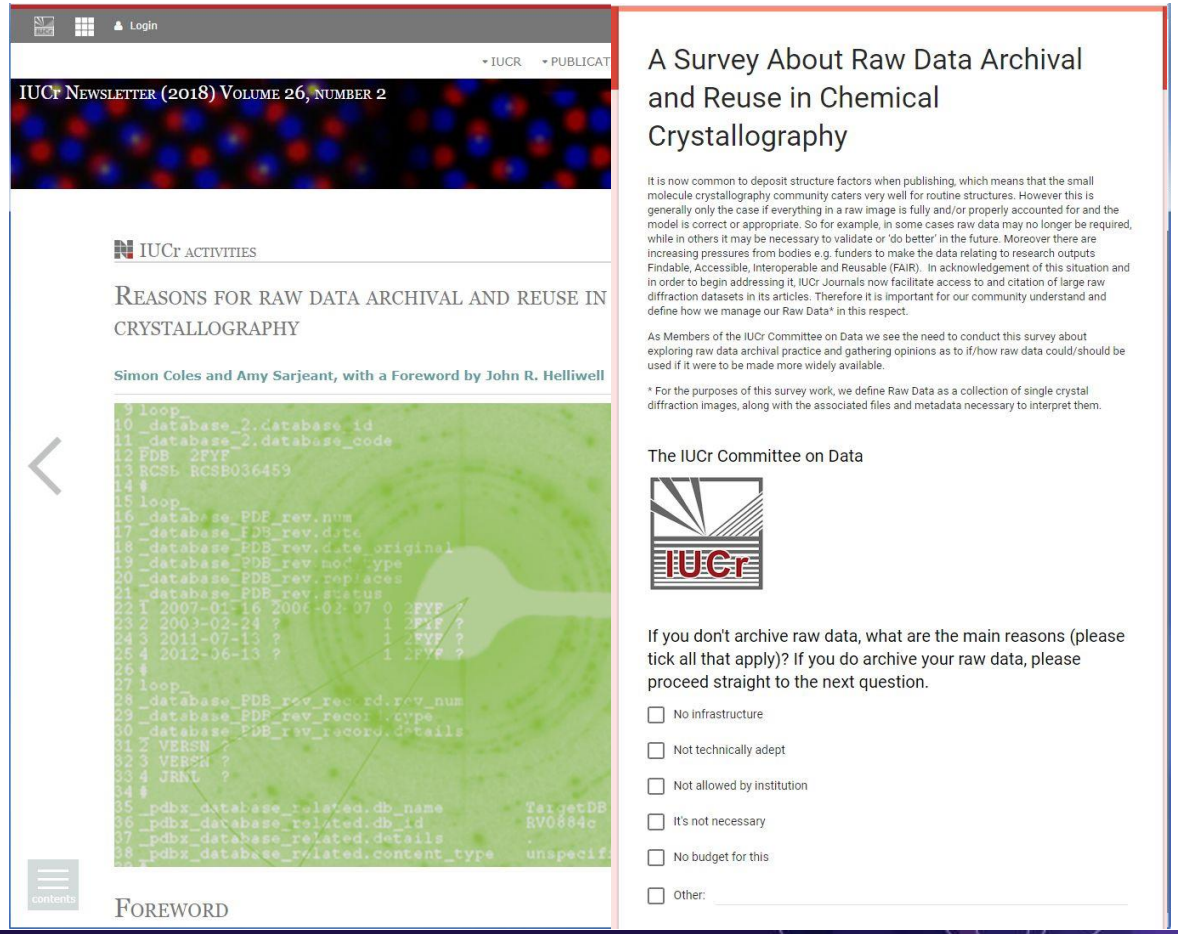


A survey by Simon Coles and Amy Sarjeant

IUCr Newsletter
Volume **26**, No. 2

Deadline 1 March 2019

Raw data archiving in chemical crystallography



A survey by Simon Coles and Amy Sarjeant

IUCr Newsletter
Volume **26**, No. 2

Deadline 1 March 2019

Current and future CommDat projects

- *checkCIF* for raw data (Loes Kroon-Batenburg, James Hester)
- Need for archiving of raw data in chemical crystallography (Simon Coles, Amy Sarjeant)
- (tentative) 2019 Workshop ‘Data Science Skills in Publishing: for authors, editors and referees’

Remember …

… the Science is in the Data

- So you need to be *able to* trust the data
- And you need to trust the data!

topical reviews

IUCrJ

ISSN 2052-2525

NEUTRON | SYNCHROTRON

The science is in the data

John R. Helliwell,[a]* Brian McMahon,[b] J. Mitchell Guss[c] and Loes M. J. Kroon-Batenburg[d]

[a]School of Chemistry, University of Manchester, Manchester M13 9PL, England, [b]International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England, [c]School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia, and [d]Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, CH 3584 Utrecht, The Netherlands. *Correspondence e-mail: john.helliwell@manchester.ac.uk

Received 18 July 2017
Accepted 24 September 2017

Edited by D. Argyriou, Ames Laboratory, USA

Keywords: raw diffraction data; sharing raw data and its reuse; open science; education; crystallographic science case studies.

Understanding published research results should be through one's own eyes and include the opportunity to work with raw diffraction data to check the various decisions made in the analyses by the original authors. Today, preserving raw diffraction data is technically and organizationally viable at a growing number of data archives, both centralized and distributed, which are empowered to register data sets and obtain a preservation descriptor, typically a 'digital object identifier'. This introduces an important role of preserving raw data, namely understanding where we fail in or could improve our analyses. Individual science area case studies in crystallography are provided.

1. Introduction

1.1. The significance of all sorts of scientific data

The meaning of the title of this article seems almost self-evident. For scientific inquiry, the 'data' are what we collect to explore nature, to test hypotheses and to suggest novel properties and mechanisms, as well as make 'findings of fact'. Yet 'data' is a very broad term. In crystallographic structure experiments, it may refer to the 'raw' data, such as diffraction images collected at the diffractometer (although even these are not truly 'raw', inasmuch as they are captured according to the electronics and mechanical properties of the detector, with whatever limitations or shortcomings are inherent to that particular device). It may also refer to the 'processed' data – for example merged structure factors – that result from calibration, reduction and other manipulation of the original images, and that constitute the material for structure solution and model refinement. The term 'data' is also used for the itemized description of the derived structural model itself (as in the coordinate sets and anisotropic displacement parameters stored in structural databases).

In all of these categories, crystallographic science resides. With raw diffraction data sets, we capture as much information as we can about the atoms of a crystal *in situ*. With the processed diffraction data sets, we retain an averaged description of the structural units in the crystal, but we may have ignored diffusely scattered intensities that contain information about disorder or large-scale correlations or we may have ignored a second crystal lattice, as in the case of pseudo-merohedral twinning. By the time we consider derived structure models, we have largely idealized a molecular structure or a 'typical' atomic environment. At each step, our level of abstraction is (usually) appropriate to the study at

OPEN ACCESS

714

Helliwell, J. R., McMahon, B., Guss, J. M. & Kroon-Batenburg, L. M. J. (2017). The science is in the data. *IUCrJ* **4**, 714-722.

ALPSP Award 2006: IUCr

Not secure | www.iucr.org/news/newsletter/volume-14/number-3/international-award-for-publishing-innovation

Login

IUCr · PUBLICATIONS · NEWS · EDUCATION · PEOPLE · RESOURCES · OUTREACH

IUCr Newsletter (2006) Volume 14, number 3

Awards and prizes

IUCr wins prestigious international award for publishing innovation

2006 AWARDS ALPSP WINNER

The IUCr has won the 2006 Award for Publishing Innovation of the Association of Learned and Professional Society Publishers (ALPSP).

The Award, for Data Exchange, Quality Assurance and Integrated Data Publication (CIF and *checkCIF*), recognizes the involvement of the IUCr in the development of the Crystallographic Information Framework and its applications, for example

- standard data definitions for crystallographic information archive and interchange
- submission format for structure report articles in crystallographic journals
- standard format for depositing supplementary structural data accompanying publications
- automated checking of the integrity and self-consistency of crystal structure models (the web checkCIF service)
- use of *checkCIF* as a peer review tool
- dissemination of crystal structural models in online publications and automated visualization.

Brian McMahon (left), IUCr R&D Officer, receives the award from Bernard Donovan, ALPSP President, at the Association's annual dinner on 14 September 2006 at the Armourers' Hall in London, UK.

The judges felt that in developing CIF and *checkCIF*, the IUCr has established an important example of data quality assurance with potential applications in other scientific, medical and indeed social sciences publishing.

The crystallographic information file (CIF) and associated data dictionaries allow the seamless transfer of information from experimental apparatus, through computation analysis, to database deposition and publication. CIF also allows the definition of quality standards for data deposition and publication and the deployment of mechanisms for checking compliance with such standards, via the checkCIF web-accessible service. The main features of the standard are its well-defined machine-readable syntax, large collection of individually defined data names and the formalism that allows automatic validation of certain attributes of data. The standard has been extended to assist with routine aspects of editorial checking and peer review, which can now take place with much increased speed and confidence.

contents

CODATA Prize 2014: Sydney R. Hall, IUCr

DATA SCIENCE JOURNAL

Reading: The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation o...

Share:

Research Papers

The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data

Authors: Sydney R. Hall, Brian McMahon

Abstract

The global application of the Crystallographic Information Framework (CIF) to the molecular structure domain has been successful over the past 20 years. It is used widely by molecular science journals and databases for submission, validation and deposition. This paper will give an overview of the CIF implementation, highlighting its particular successes and occasional failures. It will also recommend criteria for the application of an ontology-based data management system to any information domain. The paper will conclude with some details of the latest STAR data definition language and the importance of methods to the preservation of derivative data items.

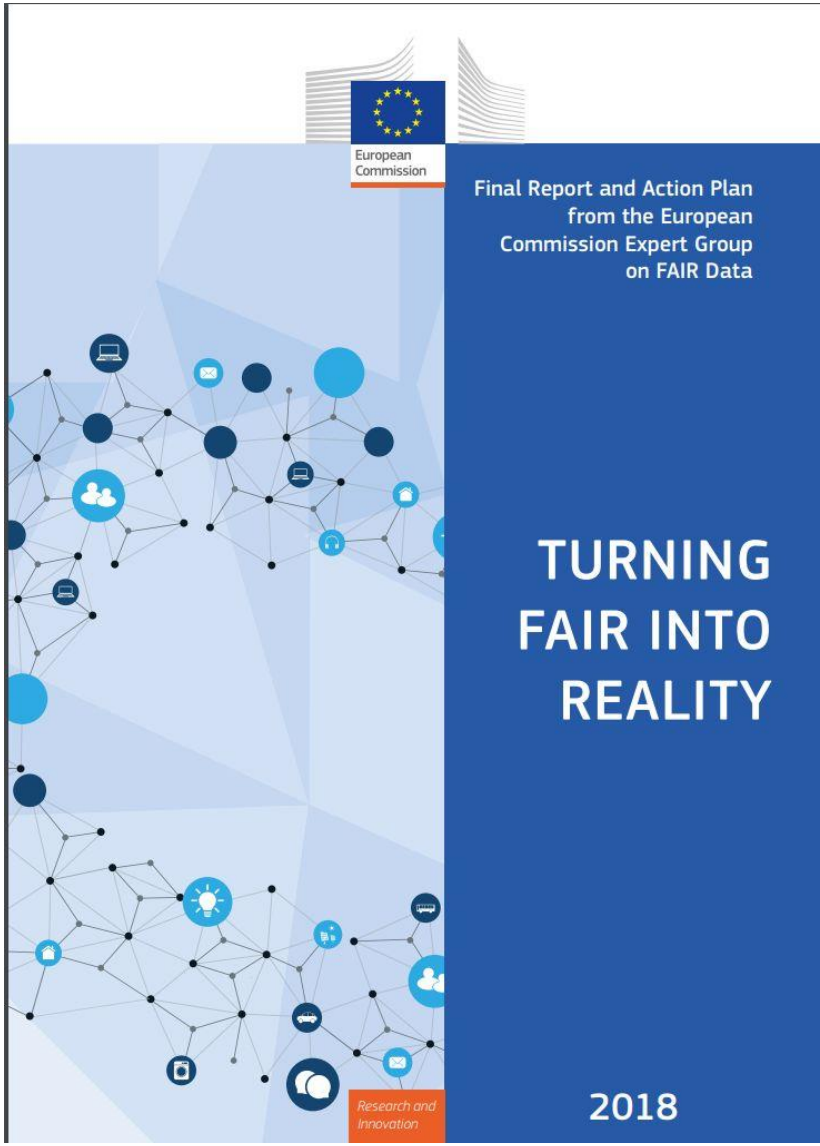
Keywords: crystallography, crystallographic information file, data exchange standard, machine readable dictionaries, ontologies

- Hall, S.R. & McMahon, B. (2016). The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Science Journal*. **15**, p.3.
DOI: http://doi.org/10.5334/dsj-2016-003

Leading the way

“The requirement from academic journals that authors provide data in support to their papers has proven to be potentially culture-changing, as has been the case in crystallography.”

“Many data standards are maintained by international scientific unions (*e.g.* the International Union of Crystallography) … As essential components of the FAIR data ecosystem there is a need for a better understanding of the business models and sustainability of the organisations that maintain specifications and standards, as well as succession plans, should current methods of maintenance and support fail.”



Thank you for your attention!