Integrating Data with Publications: Greater Interactivity and Challenges for Long-Term Preservation of the Scientific Record

Brian McMahon
International Union of Crystallography
5 Abbey Square
Chester CH1 2HU
UK
bm@iucr.org

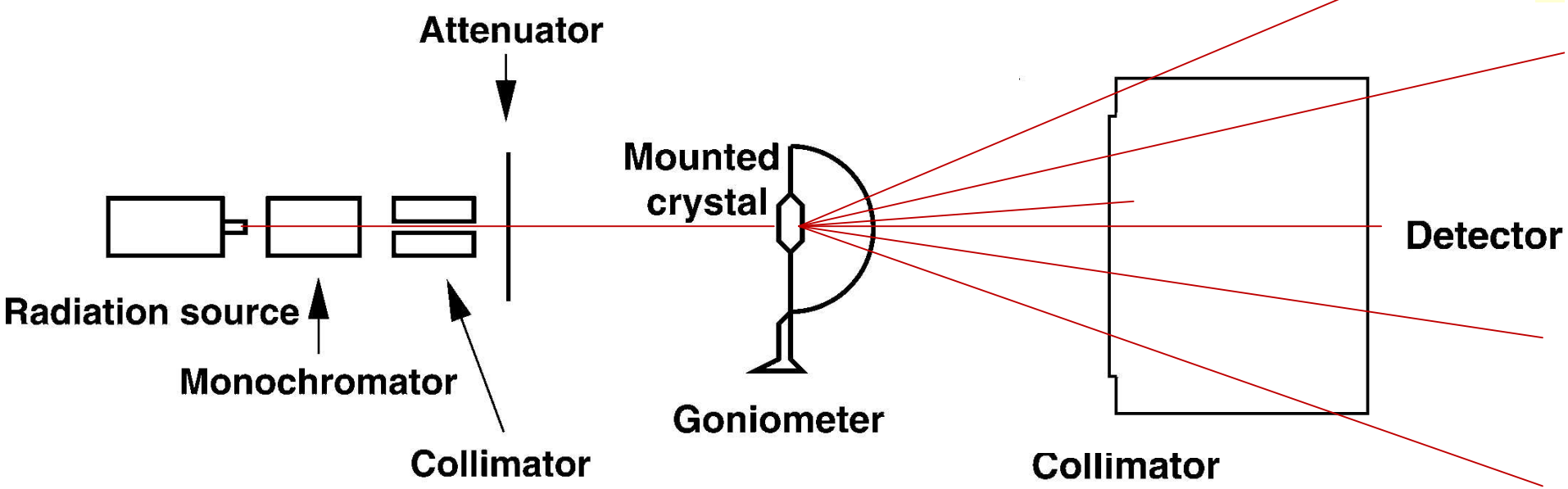
CODATA Conference 2010

Crystallography Journals Online

International Union of Crystallography

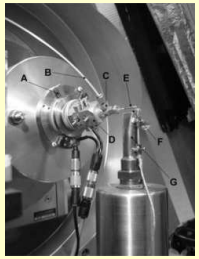
- International Scientific Union
- Publishes 8 research journals:
  - *Acta Crystallographica Section A: Foundations of Crystallography*
  - *Acta Crystallographica Section B: Structural Science*
  - *Acta Crystallographica Section C: Crystal Structure Communications*
  - *Acta Crystallographica Section D: Biological Crystallography*
  - *Acta Crystallographica Section E: Structure Reports Online*
  - *Acta Crystallographica Section F:Structural Biology and Crystallization Communications*
  - *Journal of Applied Crystallography*
  - *Journal of Synchrotron Radiation*
- Publishes major reference work *International Tables for Crystallography* (8 volumes)
- Promotes standard crystallographic data file format (CIF)

Crystallographic (X-ray diffraction) experiment

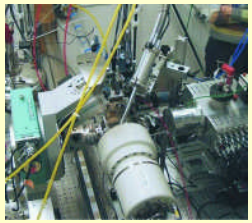


Before crystal

At crystal



After crystal



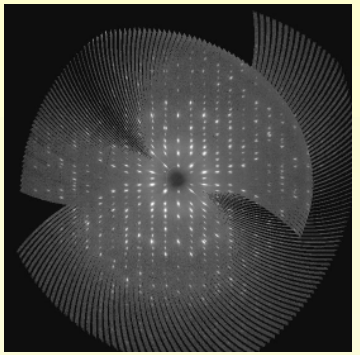
Crystallography Journals Online

Types of data relevant to publication

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 that are of secondary interest to the current focus of attention.



(1)

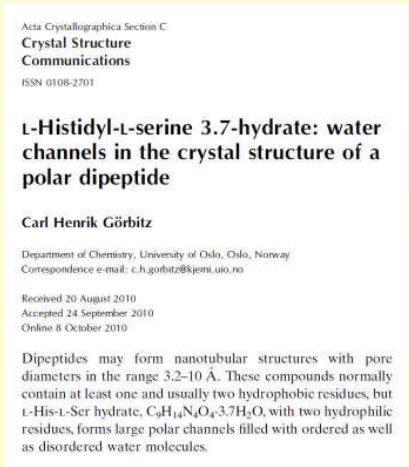
```
# h,k,l, Fc-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_refln_list_code          4
_shelx_F_calc_maximum      183.83
_exptl_crystal_F_000       1144.00
_refins_d_resolution_high  0.7705

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     39.55 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      3.84 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    546.30 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)



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

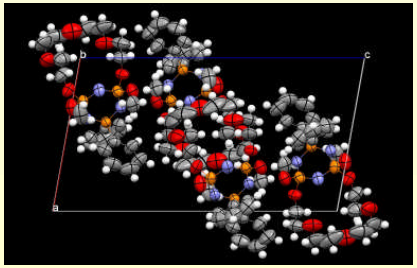
Carl Henrik Görbitz

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)



(4)

Crystallography Journals Online

Why publish data?

Some reasons:

- To enhance the reproducibility of a scientific experiment
- To verify or support the validity of deductions from an experiment
- To safeguard against error
- To safeguard against fraud
- To allow other scholars to conduct further research based on experiments already conducted
- To allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
- To provide example materials for teaching and learning
- To provide long-term preservation of experimental results and future access to them
- To permit systematic collection for comparative studies

Crystallography Journals Online

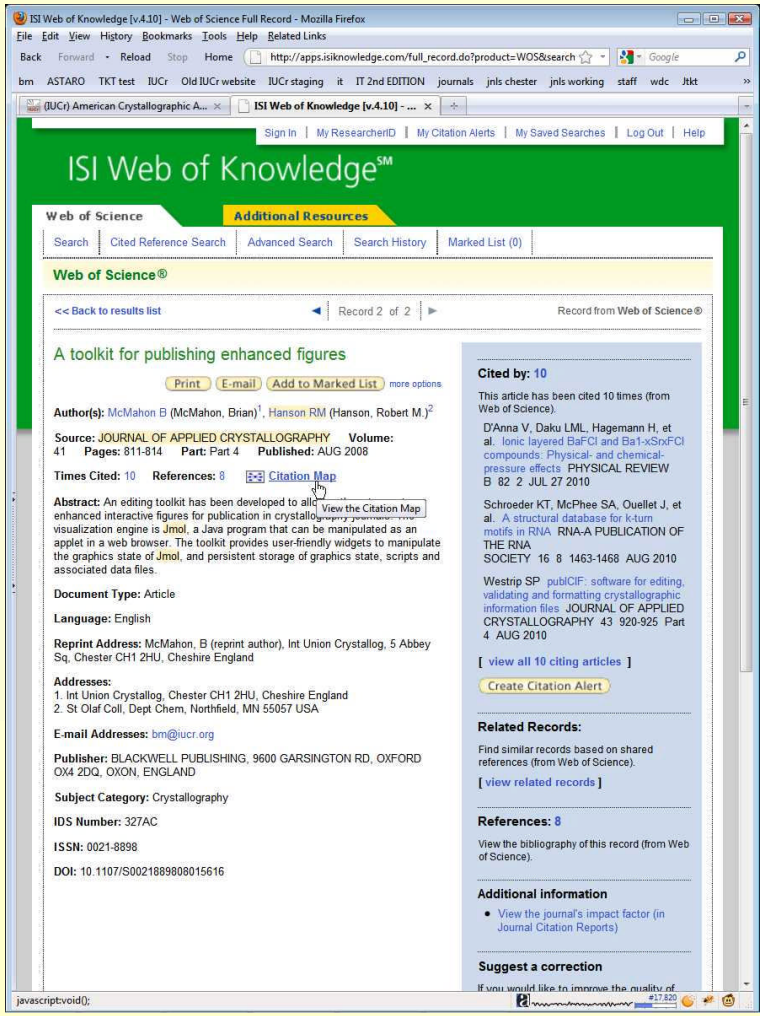
IUCr journal policy
(0) Bibliographic and linking 'metadata'

Standard propagation through publishing industry channels:

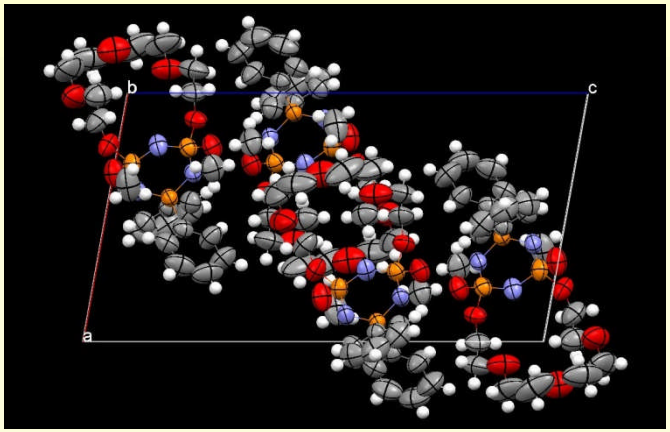
- Abstracting and indexing services
- Bibliographic databases
- Digital Object Identifiers (DOI) – registered through CrossRef
- OpenURL discovery services
- RSS feeds
- Indexing by Google Scholar, Microsoft etc.

JISC #jiscopenbib
Collaboration with Open Knowledge Foundation, Unilever Cambridge Centre for Molecular Informatics;
support from Public Library of Science, Oxford University ('Open Citation' project)

- decouple bibliographic metadata from IP issues
- pragmatic metadata harvesting from diverse schemas



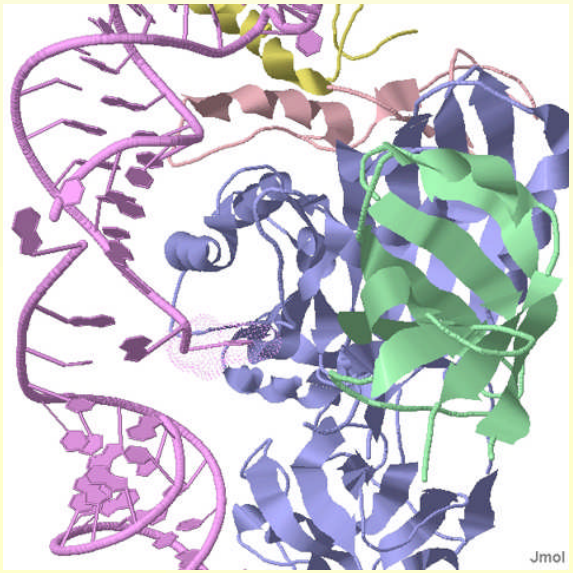
IUCr journal policy
(1) Derived data



For crystal/molecular structures with small unit cells (inorganic, metal-organic, organic)

- Atomic coordinates, anisotropic displacement parameters, molecular geometry and intermolecular contacts
- Experimental parameters, unit-cell dimensions, space group information
- Reference and modulated structure subsystems for aperiodic composite structures

must be supplied in CIF format as an integral part of article submission and are freely available for download



For biological macromolecular structures

- Atomic coordinates, anisotropic or isotropic displacement parameters, space group information, secondary structure and information about biological functionality
  must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code.
- Relevant experimental parameters, unit-cell dimensions

are required as an integral part of article submission and are published within the article

IUCr journal policy

(2) Processed experimental data

For crystal/molecular structures with small unit cell (inorganic, metal-organic, organic)

- Structure factors
- Rietveld profiles

must be supplied in CIF format as an integral part of article submission and are freely available for download. *SHELXL* instruction files are also required for validation

For biological macromolecular structures

- Structure factors

must be deposited with the Protein Data Bank before or in concert with article publication; the article will link to the PDB deposition using the PDB reference code.

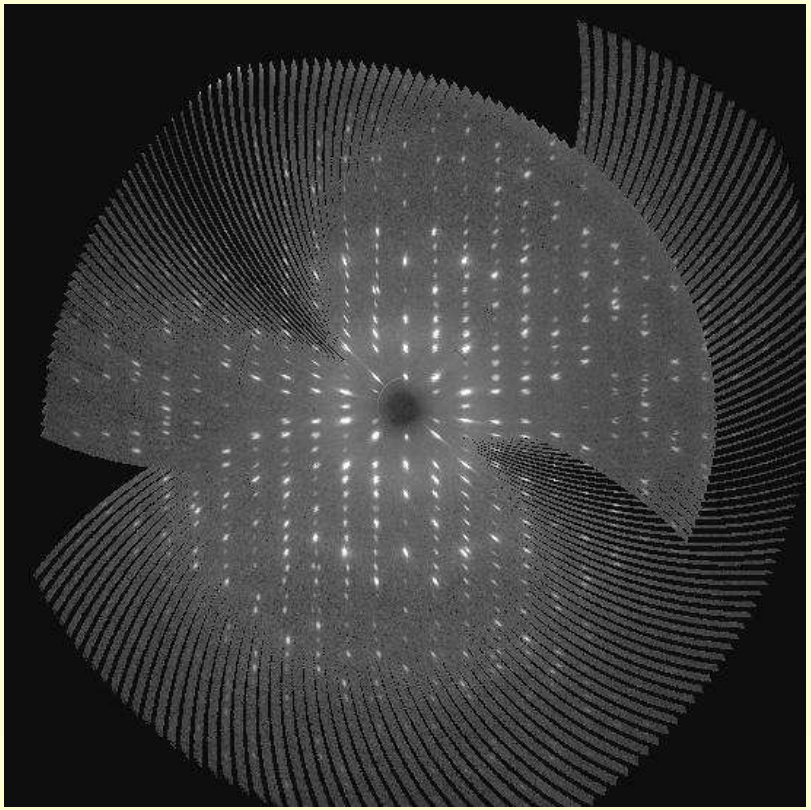
```
# h,k,l, Fc-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_refln_list_code          4
_shelx_F_calc_maximum      183.83
_exptl_crystal_F_000      1144.00
_reflns_d_resolution_high  0.7705

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      39.55 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       3.84 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     546.30 o
   2   2   0    20572.37    19502.51     520.01 o
   3   2   0     8854.88     8282.53     169.57 o
```

IUCr journal policy

(3) Primary experimental data

For crystal/molecular structures with small unit cell and for biological macromolecular structures:

IUCr journals have no current policy regarding publication of diffraction images or similar raw data entities. However, IUCr Commissions are interested in the possibility of establishing community practices for the orderly retention and referencing of such data sets, and the IUCr would like to see such data sets become part of the routine record of scientific research in the future.



- Typical size of raw data set (collection of diffraction images) ~ few Gb
- Not large enough to warrant dedicated data centres
- Too large for existing database operations (CCDC, PDB)
- Retention by individual scientists ~ 1 year
- Possibility of retention by laboratories/experimental facilities
    - o Distributed archive
    - o Requires identification/linking protocols to publications

Crystallography Journals Online

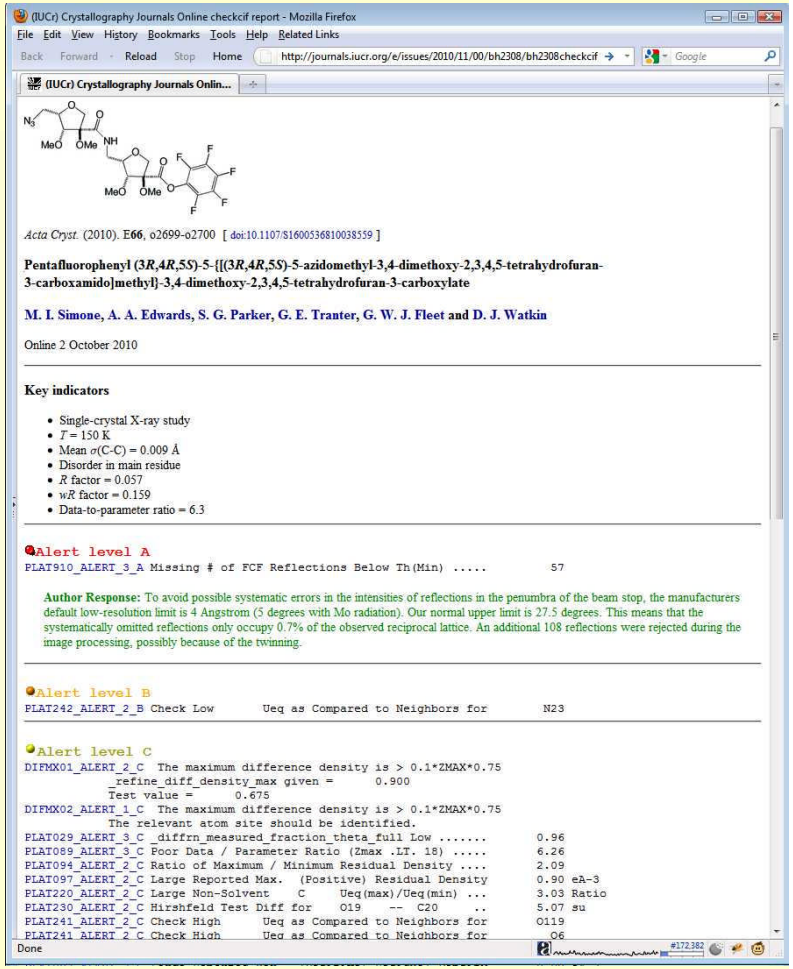
Validation of published data

For crystal/molecular structures with small unit cell (inorganic, metal-organic, organic)

- All structural models are processed upon article submission by the IUCr *checkCIF* suite and the resultant report scrutinised during the peer review process
- Structure factor files are now analysed with *checkCIF* and a report also generated for review
- *checkCIF* reports are published as supplementary documents for every *Acta Cryst. E* article
- *SHELXL* refinement instruction files are required

For biological macromolecular structures

- Structural models are annotated and curated by PDB staff during deposition; authors are consulted and given opportunities for revision if errors or anomalies are found
- Structure factor files are checked at the PDB against the deposited structure
- Data validation reports must be supplied to IUCr journals


(IUCr) Crystallography Journals Online checkcif report - Mozilla Firefox

File Edit View History Bookmarks Tools Help Related Links

Back Forward Reload Stop Home http://journals.iucr.org/e/issues/2010/11/00/bh2308/bh2308checkcif

(IUCr) Crystallography Journals Onlin...

Acta Cryst. (2010). E**66**, o2699-o2700 [doi:10.1107/S1600536810038559]

Pentafluorophenyl (3*R*,4*R*,5*S*)-5-{[(3*R*,4*R*,5*S*)-5-azidomethyl-3,4-dimethoxy-2,3,4,5-tetrahydrofuran-3-carboxamido]methyl}-3,4-dimethoxy-2,3,4,5-tetrahydrofuran-3-carboxylate

M. I. Simone, A. A. Edwards, S. G. Parker, G. E. Tranter, G. W. J. Fleet and D. J. Watkin

Online 2 October 2010

Key indicators

- Single-crystal X-ray study
- *T* = 150 K
- Mean σ(C-C) = 0.009 Å
- Disorder in main residue
- *R* factor = 0.057
- *wR* factor = 0.159
- Data-to-parameter ratio = 6.3

Alert level A

PLAT910_ALERT_3_A Missing # of FCF Reflections Below Th(Min) 57

Author Response: To avoid possible systematic errors in the intensities of reflections in the penumbra of the beam stop, the manufacturers default low-resolution limit is 4 Angstrom (5 degrees with Mo radiation). Our normal upper limit is 27.5 degrees. This means that the systematically omitted reflections only occupy 0.7% of the observed reciprocal lattice. An additional 108 reflections were rejected during the image processing, possibly because of the twinning.

Alert level B

PLAT242_ALERT_2_B Check Low Ueq as Compared to Neighbors for N23

Alert level C

DIFMX01_ALERT_2_C The maximum difference density is > 0.1*ZMAX*0.75
_refine_diff_density_max given = 0.900
Test value = 0.675

DIFMX02_ALERT_1_C The maximum difference density is > 0.1*ZMAX*0.75
The relevant atom site should be identified.

PLAT029_ALERT_3_C _diffrn_measured_fraction_theta_full Low 0.96
PLAT089_ALERT_3_C Poor Data / Parameter Ratio (Zmax .LT. 18) 6.26
PLAT094_ALERT_2_C Ratio of Maximum / Minimum Residual Density 2.09
PLAT097_ALERT_2_C Large Reported Max. (Positive) Residual Density 0.90 eA-3
PLAT220_ALERT_2_C Large Non-Solvent C Ueq(max)/Ueq(min) ... 3.03 Ratio
PLAT230_ALERT_2_C Hirshfeld Test Diff for O19 -- C20 ... 5.07 su
PLAT241_ALERT_2_C Check High Ueq as Compared to Neighbors for O119
PLAT241_ALERT_2_C Check High Ueq as Compared to Neighbors for O6

Done

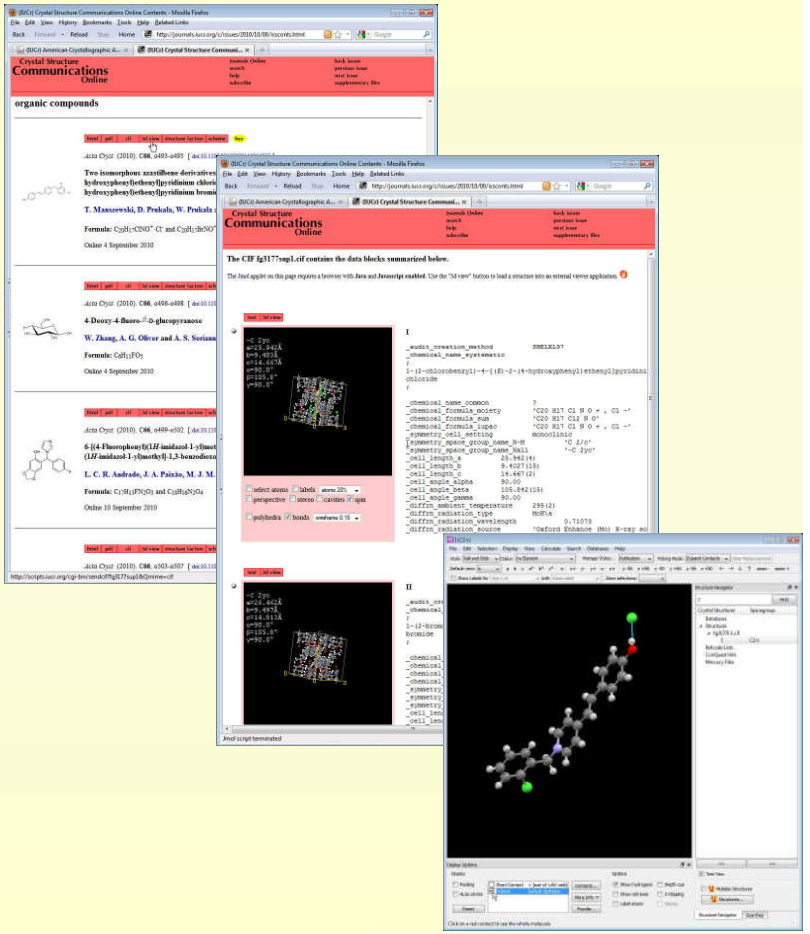
Crystallography Journals Online

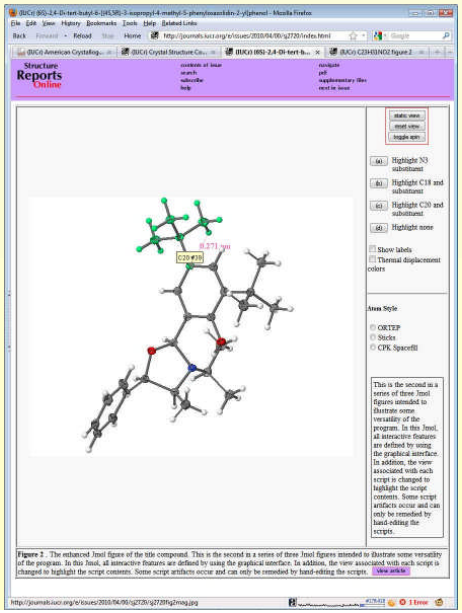
Data visualization

Three-dimensional structural models can be visualized interactively in IUCr journals in several ways:

- From journal contents lists
  - '3d view' provides *Jmol* applet (all CIFs)
  - '3d view' allows browsers to launch helper applications such as *Mercury*, *Rasmol* (multi-structure CIFs)

- Within journal articles
  - Enhanced *Jmol* figures created by authors
  - Links to PDB visual representations





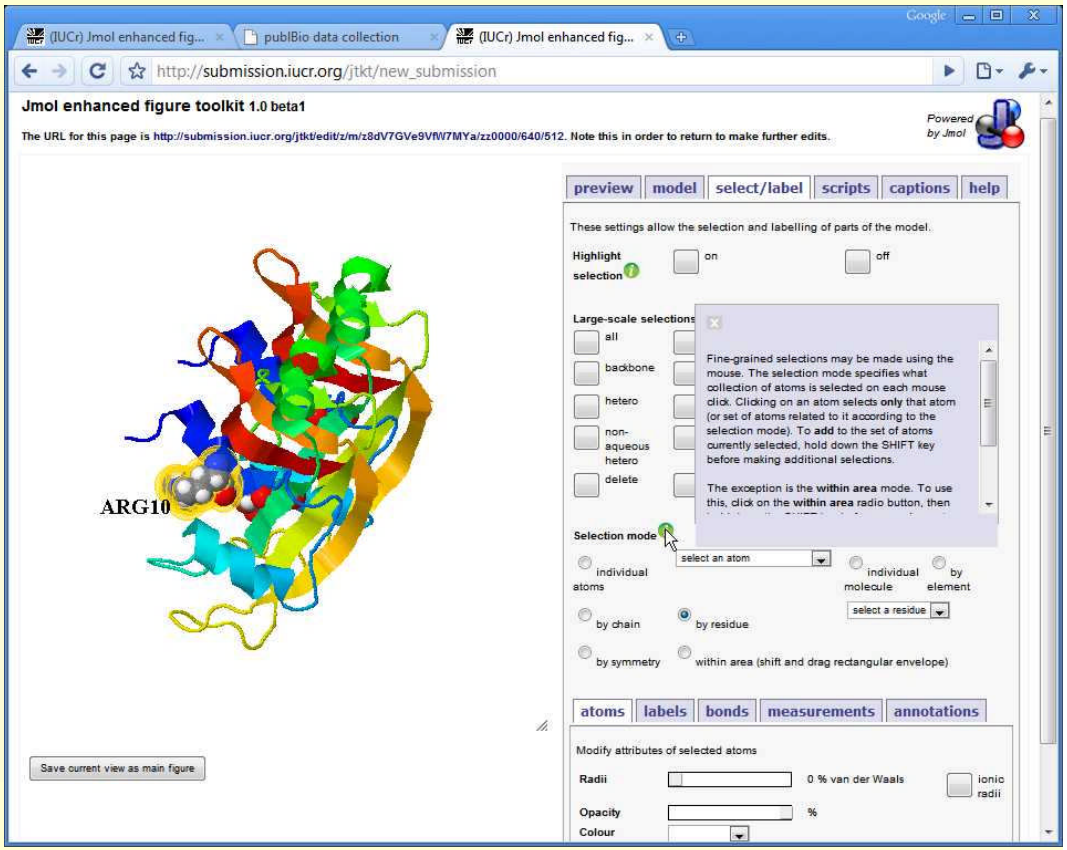
Crystallography Journals Online

IUCr enhanced figure toolkit

http://submission.iucr.org/jtkt

Create *Jmol* enhanced figures

- prior to submission of articles to IUCr journals
- during submission to IUCr journals
- from submitted CIF for structural articles
- from other CIF uploaded by authors
- from structures already published in IUCr journals
- from structures deposited in the PDB
- inorganic crystal lattices
- organic or metal-organic complexes
- biological macromolecules
- multiple views specified by author
- supports arbitrary *Jmol* scripts



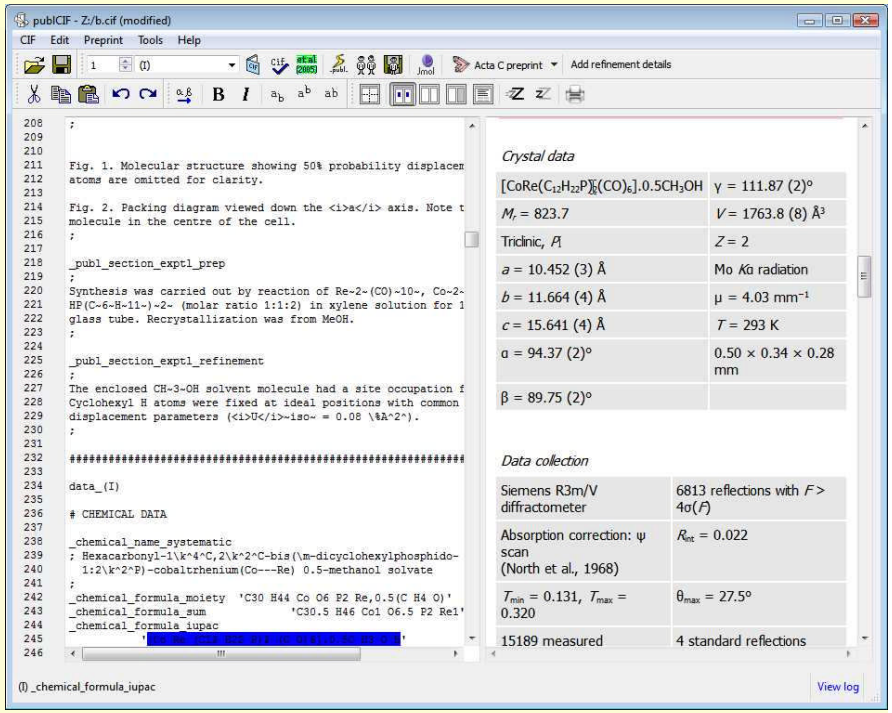
Crystallography Journals Online

Authoring tools
(1) *publCIF*

Desktop CIF publishing editor, validator and formatter for small-molecule, powder, modulated and incommensurate structure CIFs
http://journals.iucr.org/services/cif/publcif/

Available for Linux, Microsoft Windows, MacOS X
publCIF takes a CIF and prepares a formatted paper (Preprint) in the style of *Acta Crystallographica Sections C* and *E*.
CIF and Preprint are presented side-by-side and are both editable. Changes made to one are applied to the other as the user types.
Starting with a CIF resulting from a structure refinement, use *publCIF* to:

- add data items required for publication, using simple wizards that check input and access both dictionary information and data used in previous CIFs
- prepare standard and customized geometry tables using simple spread sheets
- write paper using a word-processing environment
- check CIF for both syntax and completeness, with facility to load directly into online *checkCIF*
- print or export Preprint of a paper
- check references
- store or retrieve author details from a local database



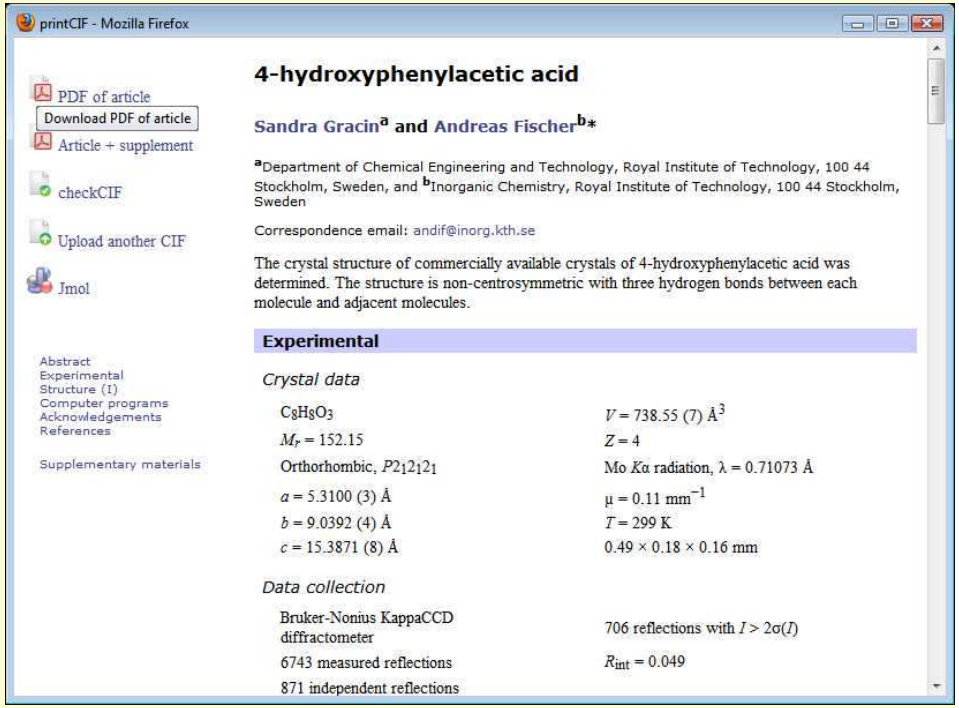
Crystallography Journals Online

Authoring tools
(2) *printCIF*

Online CIF publishing validator and formatter for small-molecule, powder, modulated and incommensurate structure CIFs
http://publcif.iucr.org/services/tools/printcif.php

printCIF takes a CIF and prepares a formatted paper (Preprint) in the style of *Acta Crystallographica Sections C* and *E*. Starting with a CIF resulting from a structure refinement, use *printCIF* to:

- create HTML version of article and supplementary information with live *Jmol* visualization of molecular geometry
- create PDF version of article
- create PDF version of article and supplementary document
- create PDF version of supplementary material
- run online *checkCIF* validation
- visualize structure in *Jmol*



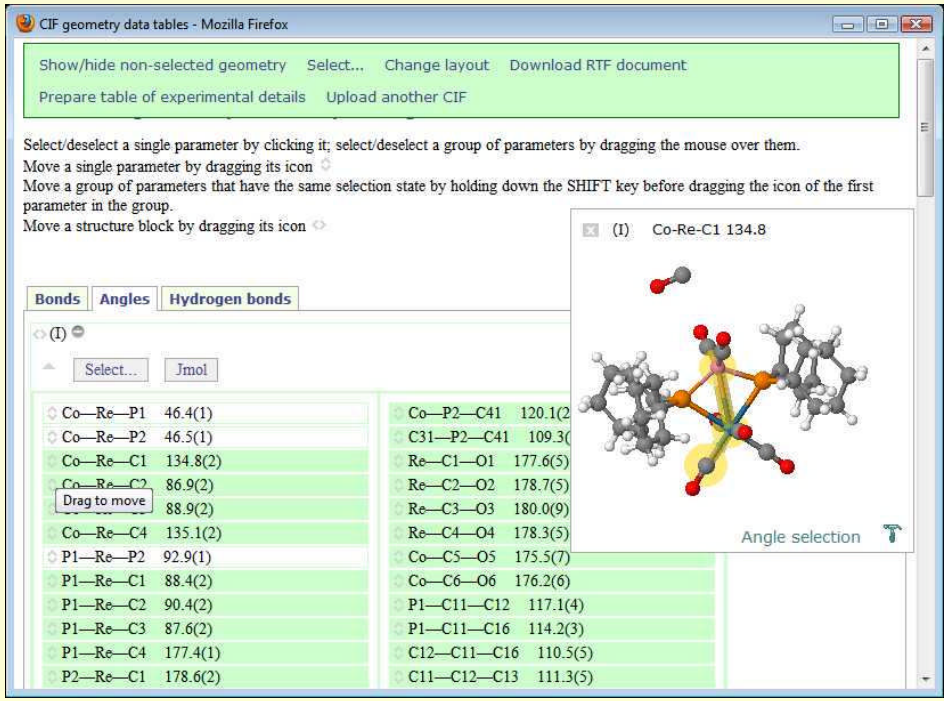
Crystallography Journals Online

Authoring tools
(3) experimental tables

Online service for formatting complex geometry and experimental tables for small-molecule, powder, modulated and incommensurate structure CIFs
http://publcif.iucr.org/services/tools/

Takes a CIF and prepares formatted experimental or geometry tables (most useful for *Acta Crystallographica Section B* but also suitable for other IUCr journals).
Starting with a CIF resulting from a structure refinement, use this service to:

- create HTML or RTF versions of experimental and geometry tables
- reorder and relabel experimental tables
- show only information from selected data blocks
- add data items not required by the default request list
- reorder and suppress entries in molecular geometry tables
- reformat geometry tables
- visualize geometry in *Jmol*
- save CIF modified in accordance with reformatted presentation



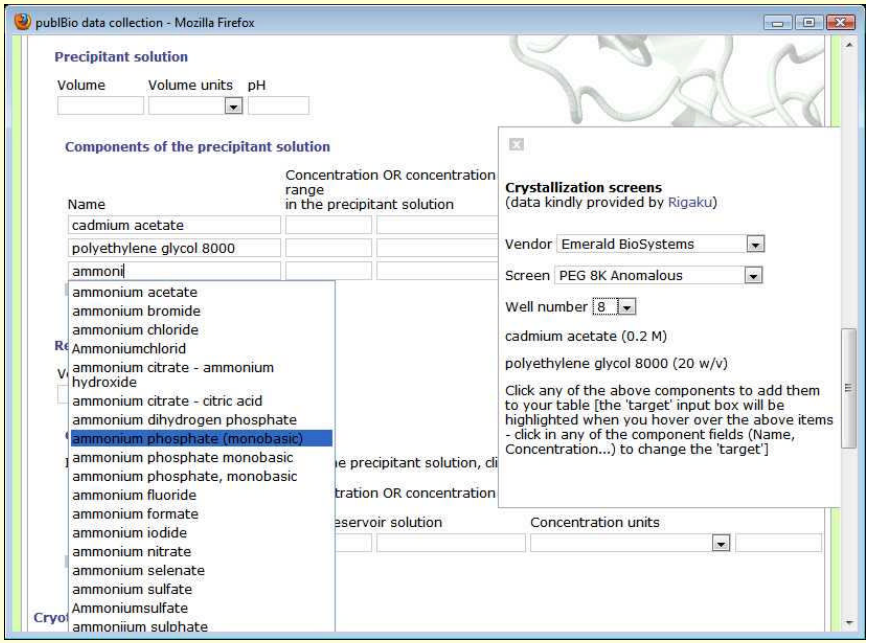
Crystallography Journals Online

Authoring tools
(4) *publBio*

Online service for authoring and editing structural biology or crystallization communications using macromolecular CIFs (mmCIF)
http://publbio.iucr.org

Provides a personal web space for creating, editing, managing and submitting these categories of articles to *Acta Crystallographica Sections D* and *F*.
Starting with a local mmCIF, a PDB reference, or an empty template, use this service to:

- add data items required for publication, using a structured forms-based graphical user interface that checks input and accesses both dictionary information and data used in previous CIFs
- write paper using WYSIWYG and predictive type-ahead forms
- check and format citations and literature references
- incorporate static and *Jmol* interactive figures
- create HTML, OpenOffice, RTF or PDF versions of an article
- lookup common crystallization solution components database
- maintain a local database of author information
- lookup references from Medline or Crystallography Journals Online



Crystallography Journals Online

The future

Not all crystallography journals accept structural data for deposit, and IUCr journals do not (at present) accept primary data sets. Ensuring that data are retained in the permanent record of science will require:

- increased willingness by journals to deposit structural data and structure factors
- increased willingness of structural databases to store and curate structural data, structure factors and perhaps primary data
- increased collaboration between journals and structural databases that archive data sets
- assignment of persistent identifiers to unpublished data held in
    - domain-specific repositories (*e.g.* eCrystals at Southampton)
    - institutional repositories
    - synchrotron and other experimental facilities
    - image data stores (*e.g.* Atlas, Rutherford Appleton Laboratories)
- community codes of practice for storing, managing, archiving and accessing research data sets
- standard 'compound document' descriptions to link data and publications

Crystallography Journals Online

References and further reading

Structure validation: P. R. Strickland, M. A. Hoyland and B. McMahon, Small-molecule crystal structure publication using CIF. *International Tables for Crystallography* (2006). Vol. G, Chapter 5.7, pp. 557–569.

Enhanced figure toolkit: B. McMahon and R. M. Hanson, A toolkit for publishing enhanced figures. *Journal of Applied Crystallography* (2008), **41**, 811-814.

publCIF: S. P. Westrip, publCIF: a CIF publishing editor, validator and formatter. *Journal of Applied Crystallography* (2010), submitted.

Research data management policy: UK Research Data Service, *The data imperative: Managing the UK's research data for future use* (2009). Available from http://www.ukrds.ac.uk/resources/download/id/14

Research data management policy: e-IRG Data Management Task Force. *Report on Data Management* (2009). Available from http://www.e-irg.eu/images/stories/publ/task_force_reports/dmtfjointreport.pdf

Open data principles: P. Murray-Rust, C. Neylon, R. Pollock and J. Wilbanks. *Panton Principles for Open Data in Science* (2009). Available from http://pantonprinciples.org/

Persistent identifiers for data: DataCite (2009). http://www.datacite.org/

Crystallography Journals Online