The element of trust: validating and valuing crystallographic data

Brian McMahon¹, John R. Helliwell² and James R. Hester³

¹International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK. E-mail: bm@iucr.org
²School of Chemistry, University of Manchester, Manchester M13 9PL, England
³Australian Nuclear Science and Technology Organisation, New Illawarra Road, Lucas Heights, NSW 2234, Australia
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).
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-Benzothien-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)


**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 manufacturers’ 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, single-crystal diffractometers and produce output in CIF or SCF5 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)

_data_atom_site_aniso_label_

_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

_data_atom_site_aniso_type_symbol_

_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

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: 'Angstroms squared' ('\( a \)'), 'pm' (picometres squared /10000.), 'nm' (nanometres squared /10000).

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

Experiment (synchrotron or laboratory) → Data reduction → Structure solution and refinement (laboratory) → IUCr journals

Chemistry databases (CCDC) → Other journals → Biological structure databases (PDB)

- Raw experimental data
- Reduced/processed data
- Derived data

- retained by scientist
- archived at facility (~6 months)
- deposited
- published/disseminated
- validated
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

<table>
<thead>
<tr>
<th>Resource</th>
<th>Category Type</th>
<th>Language Schema</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIF core dictionary</td>
<td>IUCr standard dictionary</td>
<td>DDL1 ftp://ftp.iucr.org/pub/dll_core_dic</td>
<td>crystal structure; molecular structure; chemical structure; symmetry; crystal morphology; sample preparation; laboratory apparatus; experimental processed data; single-crystal diffraction; chemical composition; author, principal investigator, experimenter etc.</td>
</tr>
<tr>
<td>CIF restraints dictionary</td>
<td>IUCr standard dictionary</td>
<td>DDL1 ftp://ftp.iucr.org/pub/dll_core_dic</td>
<td>crystal structure; molecular structure; relaxation; experimental technique</td>
</tr>
<tr>
<td>CIF powder dictionary</td>
<td>IUCr standard dictionary</td>
<td>DDL1 ftp://ftp.iucr.org/pub/dll_core_dic</td>
<td>crystal structure; molecular structure; chemical analysis; experimental processed data; powder diffraction; chemical composition</td>
</tr>
<tr>
<td>CIF modulated and composite structures dictionary</td>
<td>IUCr standard dictionary</td>
<td>DDL1 ftp://ftp.iucr.org/pub/dll_core_dic</td>
<td>crystal structure; molecular structure; chemical composition</td>
</tr>
<tr>
<td>CIF electron dictionary</td>
<td>IUCr standard dictionary</td>
<td>DDL1 ftp://ftp.iucr.org/pub/dll_core_dic</td>
<td>crystal structure; molecular structure; single-crystal diffraction; chemical composition</td>
</tr>
</tbody>
</table>
checkCIF

• Began as IUCr journals in-house suite of validation/consistency checks
Checking for completeness

Datablock: I

Checking for completeness
Checking for completeness

### Datablock: I

<table>
<thead>
<tr>
<th>Property</th>
<th>Calculated</th>
<th>Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>739.53</td>
<td>739.53</td>
</tr>
<tr>
<td>Space group</td>
<td>P 21 11 21</td>
<td>P 21 11 21</td>
</tr>
<tr>
<td>Monoclinic</td>
<td></td>
<td>P 21 11 21</td>
</tr>
<tr>
<td>Cell parameters</td>
<td></td>
<td>P 21 11 21</td>
</tr>
</tbody>
</table>

**Correction Method:** Not given

**Data Completeness**

- 

**Errors:**

- ALERT Level A
  - SYMM001_ALERT_1_4: Symmetry cell setting is missing. The cell setting should be one of the following:
    - triclinic
    - monoclinic
    - orthorhombic
    - tetragonal
    - rhombohedral
    - trigonal
    - hexagonal
    - cubic

- ALERT Level B
  - PLAT010_ALERT_1_4: Missing different ambient temperature. Datum ... Please Add
  - PLAT010_ALERT_4_4: Cell Data Missing, Incomplete or Out-of-Order. Please Check
  - PLAT005_ALERT_5_0: No Embedded Refinement Details Found In the CIF. Please Do!
  - PLAT010_ALERT_1_0: The Reported Crystal System is Inconsistent with P212121. 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
A service of the
International Union of Crystallography

checkCIF reports on the consistency and integrity of crystal structure determinations reported in CIF format.

Please upload your CIF using the form below.

File name: 
Choose File | No file chosen

Select form of checkCIF report
- HTML
- PDF
- PDF (recommended for CIFs that might take a long time to check)

Select validation type
- Full validation of CIF and structure factors
- Full IUCr publication validation of CIF and structure factors
- Validation of CIF only (no structure factors)

Output Validation Response Form
- Level A alerts only
- Level A and B alerts
- Level A, B and C alerts
- None

Send CIF for checking

Information about this version of checkCIF ...

Useful links
- Prepublication check for submissions to IUCr journals
- Details of checkCIF/PLATON tests
- CIF dictionary
- Download CIF editor (pubCIF) from the IUCr
- Download CIF editor (enCIFer) from the CCDC

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.
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
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

• 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)
DDDG publications

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](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](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 About Raw Data Archival and Reuse in Chemical Crystallography

It is now common to deposit structure factors when publishing, which means that the small molecule crystallography community cares very well for routine structures. However, this is generally only the case if everything in a raw image is fully and/or properly accounted for and the model is correct or appropriate. So for example, in some cases raw data may no longer be required, while in others it may be necessary to validate or do better in the future. Moreover, there are increasing demands for raw data. Thus, it is important to make the data relating to research outputs Findable, Accessible, Interoperable and Reusable (FAIR). In acknowledgement of this situation and in order to begin addressing it, IUCr Journals now facilitate access to and creation of large raw diffraction datasets in its articles. Therefore it is important for our community understand and define how we manage our raw data in this regard.

As members of the IUCr Committee on Data we see the need to conduct this survey about exploring raw data archiving practice and gather opinions on if/how raw data could be used if were to be made more widely available.

For the purposes of this survey work, we define Raw Data as a collection of single crystal diffraction images, along with the associated files and metadata necessary to interpret them.

The IUCr Committee on Data

If you don't archive raw data, what are the main reasons (please tick all that apply)? If you do archive your raw data, please proceed straight to the next question.

- No infrastructure
- Not technically adept
- Not allowed by institution
- It's not necessary
- No budget for this
- Other

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!

ALPSP Award 2006: IUCr

IUCr WINS PRESTIGIOUS INTERNATIONAL AWARD FOR PUBLISHING INNOVATION

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.

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.
CODATA Prize 2014: Sydney R. Hall, IUCr

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

  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!