

Data Exchange, Quality Assurance and Integrated Data Publication (CIF and *checkCIF*)

IUCr submission for the ALPSP Award for Publishing Innovation 2006



Crystallography
Journals
Online

Summary of IUCr journal publishing activities

The journals of the International Union of Crystallography are produced by the IUCr in Chester and published by Blackwell Munksgaard. The print editions of the journals are distributed by Blackwell Publishing in Oxford, while electronic editions of all IUCr journals are available *via* **Crystallography Journals Online** (<http://journals.iucr.org>).

The IUCr has published journals since 1948. Eight titles are currently published:

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 Online

Journal of Applied Crystallography

Journal of Synchrotron Radiation

Administrative Office: The Executive Secretary, IUCr, 2 Abbey Square, Chester CH1 2HU

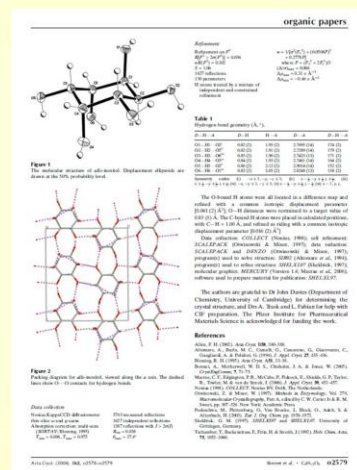
Publishing Office: The Managing Editor, IUCr, 5 Abbey Square, Chester CH1 2HU

Editorial and administrative staff (full -time equivalent): 16

Research and Development staff: 4

Crystal Structure reports - data-rich scientific articles

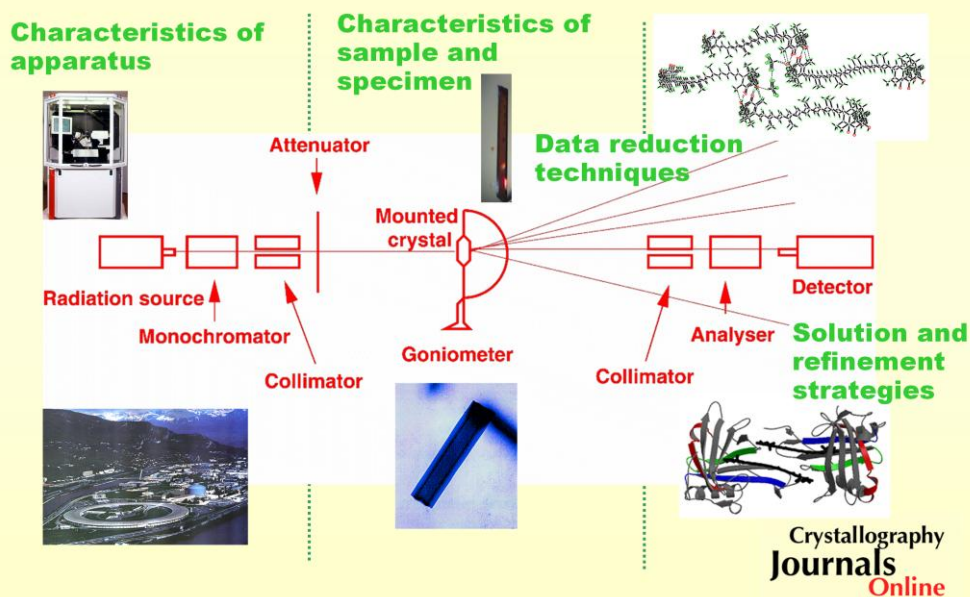
- 3-d positional coordinates
- Atomic motions
- Molecular geometry
- Chemical bonding
- Crystal packing
- Chemical behaviour arising from structure
- Two dedicated IUCr journals:
Acta Cryst. C, E
- Important part of scientific discussion in many other titles:
Acta Cryst. B, D, F



**Crystallography
Journals
Online**

Certain types of primary research literature comprise detailed discussion of the results of well-defined experiments to investigate particular properties or types of matter. Crystal structure analysis is one such. The highly ordered packing of atoms or molecules in the solid state within regular crystal lattices allows probing by experimental techniques such as X-ray diffraction. From the scattering data gathered in such an experiment, one may deduce much information about the nature and molecular structure of the components of the crystal, and this information is often sufficiently interesting to justify a new scientific publication. Among the journals of the International Union of Crystallography (IUCr), two titles contain almost exclusively such structure reports (*Acta Crystallographica Section C: Crystal Structure Communications*; *Acta Crystallographica Section E: Structure Reports Online*). They also form a routine component of longer research articles describing the chemical or physical properties arising from structures determined in this way, and so may appear in any crystallographic journal. Within the IUCr stable, they are most often found as components of articles in *Acta Crystallographica Sections B: Structural Science*; *D: Biological Crystallography*; and *F: Structural Biology and Crystallization Communications Online*.

Consistent data pipeline



We have already alluded to the variety of data storage media in use today and historically. Crystal structures can be determined by a variety of different physical techniques, using equipment varying in scale from benchtop apparatus to giant national synchrotron radiation facilities. Data are captured using different detectors (electronic counters, CCDs, photographic film *etc.*) and the different classes of crystalline material (elements, inorganic or organic compounds, polymers...) have very different properties and offer different challenges in their analysis. Nevertheless, the underlying experimental procedures can be mapped to a common procedural model, so that there is a consistent pipeline of data that need to be collected, reduced, analysed and propagated during the course of a crystal structure determination. This consistency makes it realistic to design a common data exchange standard that encompasses all the underlying variations in experimental technique.



Data exchange standard

- Crystallographic Information File (CIF)
- Machine-readable (but also human readable)
- Well-defined syntax
- Large dictionary of standard data names
- Machine-readable description of data types, value constraints, dependencies

Crystallography
Journals
Online

Such a data exchange standard has been developed by the IUCr and is now widely used in many areas of crystallography and related structural science. The original small-molecule Crystallographic Information File [Hall, S.R., Allen, F.H. & Brown, I.D. (1991), *Acta Cryst. A* **47**, 655-685] has now grown into the Crystallographic Information Framework, which includes specifications for powder diffraction, modulated and composite structures, electron density, biological macromolecules, images and symmetry. Related standards are also being developed in such areas as nuclear magnetic resonance spectroscopic determinations of macromolecular structure, protein production, and cryoelectron microscopy. The standard is now fully documented in *International Tables for Crystallography, Vol. G: Definition and exchange of crystallographic data* [Hall, S. R. & McMahon, B. (2005). Dordrecht: Springer.] The main features of the standard are its well-defined machine-readable syntax, its very large collection of individually defined data names (over 3300 distinct items, rising to 4500 if the related standards are included), and the formalism that allows automatic validation of certain attributes of data.

Examples of CIF

Data file

```
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
```

Dictionary entry

```
data_chemical_formula_weight
  _name          '_chemical_formula_weight'
  _category       chemical_formula
  _type           numb
  _enumeration_range 1.0:
  _units          Da
  _units_detail   'daltons'
  _definition
; Formula mass in daltons. This mass should
correspond to the formulae given under
  _chemical_formula_structural, *_iupac,
*_moiety or *_sum and, together with the Z value
and cell parameters, should yield the density
given as _exptl_crystal_density_diffn.
```

Crystallography
Journals
Online

The example on the left is an extract from a CIF data file, demonstrating that it can include numeric and textual data. The meaning of the data can be deduced by inspection by a knowledgeable human reader even independently of the associated data-name 'dictionaries'. Protocols exist for introducing new data names in an orderly way, ensuring the extensibility of the standard, and for indicating unknown or absent data values. The syntax is simple, quite a free format, but with sufficient regularity that it is easily machine-parsed. If the standard were being developed afresh nowadays, it would probably have an underlying XML representation. However, it was drafted before XML was invented, and has the benefit (for scientific computation) of being a format that can be handled even by Fortran programs with relatively little difficulty. It can easily be converted to XML for interoperability with non-crystallographic applications. The example on the right is one among the thousands of individual definitions in the machine-readable data-name dictionaries. It is expressed in the same syntax as the data files, so that the same parsing engine can be used to manipulate both data files and the accompanying dictionaries.



Adoption of CIF for publication

- New data names
 - Bibliographic metadata
 - Textual comment
- Technique for extracting, sorting and merging data items as required for a structure report
- Conversion CIF→TeX (later SGML) for typesetting
- CIF becomes machine-readable submission medium
- Checking presence of required data items

Crystallography
Journals
Online

Although CIF was introduced for data pipelining, and was initially seen as possibly a candidate file format for depositing supplementary data sets with journals, it was recognised by the IUCr journal production staff that with a few modest extensions it could be used as a significant component in the publication workflow. The extensions comprised: the specification of some additional data names indicating information about authors and bibliographic information; the introduction of new data items containing the scientific comment that would appear in a structure report article; a procedure for extracting the standard data that were always reported, reordering and merging these with the commentary text; and a mechanism to convert the resultant collection of data items to a typeset article. As far back as 1990, articles could be proofed on the fly from CIF data files, using TeX as an intermediary composition engine. Between 1997 and 1999 a number of shorter articles were published online only, *in raw CIF format*. However, with the launch in 1999 of the IUCr's online journals platform **Crystallography Journals Online**, all subsequent articles were made available in the HTML and PDF formats that are more familiar to regular web users. Nowadays the CIF content is transformed to SGML for complete integration in the journal production workflow. With the implementation of these extensions, CIF itself became in 1996 the standard medium for submission of structure report articles to *Acta Cryst. Sections C* and *E*, and remains so to this day. From the outset, it was recognised that the structured format of these files made it possible to devise computerised procedures for checking that the expected content of a structure report was present and complete, so that routine aspects of editorial checking could start to be automated.



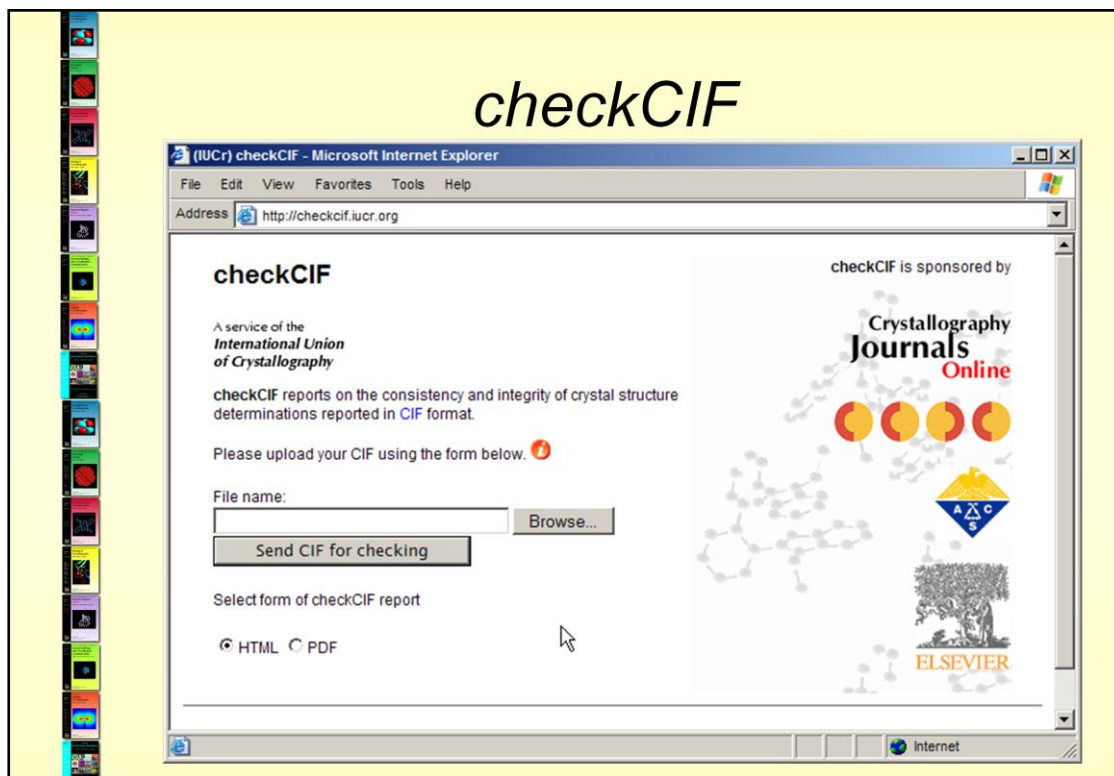
Extending validation

- Check internal consistency of data dependencies (CIF dictionary)
- Check scientific reasonableness of model
- Check completeness of experimental metadata
- Check quality of derived structural model

Crystallography
Journals
Online

The next development was to use the same ability to extract specific items of data to check the self-consistency of the numerical values reported in the file. Initially the validation tests were of the integrity of the data (that the data present had the correct type, were not outside physically reasonable bounds of numerical value *etc.*). In time, the details of the reported structural model could actually be compared against a knowledge base of chemical parameters, so that warnings could be generated if the model departed significantly from chemical precedent. Subsequently, editorial policies were introduced to check that the experimental *metadata* (information about the measurement strategy, reduction methodology and solution and refinement procedures) indicated that the derived results were of acceptable quality. From this point onwards, various criteria could be set out for judging objectively the quality of the experimental and derived data.

checkCIF



This validation analysis was developed over a decade by the academic and technical editors of the journals, working in collaboration with research scientists to develop the necessary analytic software. The result is the web service *checkCIF* (<http://checkcif.iucr.org>), sponsored by the IUCr and several publishing and database partners, that will subject any structural CIF data set to a large collection of algorithmic tests (346 distinct tests are currently listed on the IUCr web site at <http://journals.iucr.org/services/cif/datavalidation.html>).



Uses of *checkCIF*

- Pre-submission checks by author
- Objective assessment of derived data for use in peer review
- Detection of gross errors
- Quality assurance in non-specialist publications (with some crystallographic judgement)
- Analysis available to readers (*Acta E*)

Crystallography
Journals
Online

The *checkCIF* service is an integral part of the submission and review cycle for IUCr journals. Authors are encouraged to use it prior to formal submission to ensure that their files are free from syntax errors, and do not exhibit gross errors. Upon submission, the software automatically generates a report on the quality and consistency of the structure that is passed to the referee as an integral part of the peer review process. (While warnings raised by *checkCIF* are helpful in allowing the reviewer to assess the overall scientific argument presented in an article, the community still insists on the acceptance or rejection of articles being a matter of considered human judgement.) Other publishing partners in the *checkCIF* project do not rely on *checkCIF* output to the same extent, but they do use it to assist their peer reviewing procedures to the extent that they consider appropriate. One rather specific use of *checkCIF* in IUCr journals is the provision on the journal web site of a *checkCIF* validation report as a supplementary document supporting every article published in *Acta Crystallographica Section E*.

Semi-automated peer review

Pre-submission checks

File syntax

Data validation criteria for publication

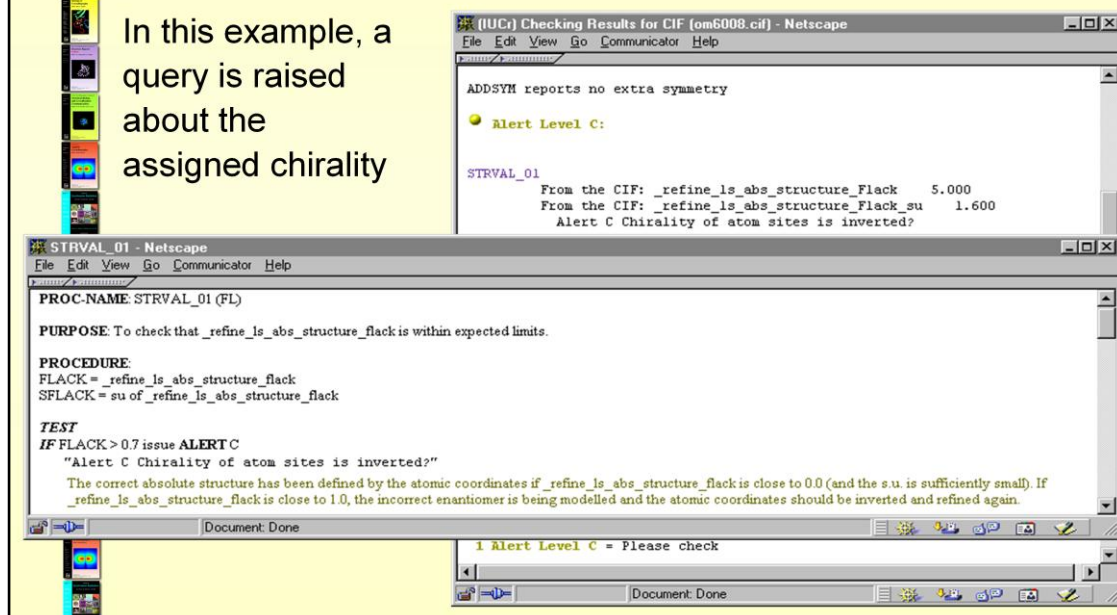
Thorough analysis and display of structure

The screenshot shows a Netscape browser window titled "(IUCr) CIF Checking Submission Form - Netscape". The page content includes a welcome message for checkCIF, a description of the service, and a form for file submission. The "File Name:" field contains "G:\example\example6008.cif" and a "Browse..." button is next to it. Below this is a "Send file for checking" button. The form also has a section for selecting options, with three radio buttons: "Basic syntax checks (checks the syntax of the file only)", "Publication checks (uses the data validation criteria for IUCr journals)" (which is selected), and "Ellipsoid plot, listing of geometry and symmetry". A caution note states: "(Caution: may take several minutes and generate listings in excess of 1Mb)". At the bottom, there are two checkboxes: "Plot atomic displacement ellipsoid diagram" (checked) and "Intermolecular contacts and H-Bonds" (unchecked). The browser's status bar at the bottom shows "Document: Done".

This illustrates how checkCIF is used in the submission/review cycle for IUCr journals. The prospective author uploads a CIF *via* a web form. The *checkCIF* software first checks the file syntax - if there are any errors, it is **immediately** returned for correction. The standard mode for checking validates against the series of algorithms published on the IUCr web pages. Optionally, the author may request a detailed listing of an analysis of the structure. This includes complete intra- and intermolecular geometry, bond types, displacement ellipsoids, structural voids, least-squares planes; and on occasion suggests peculiar features of the structure that have not before been obvious to the author.

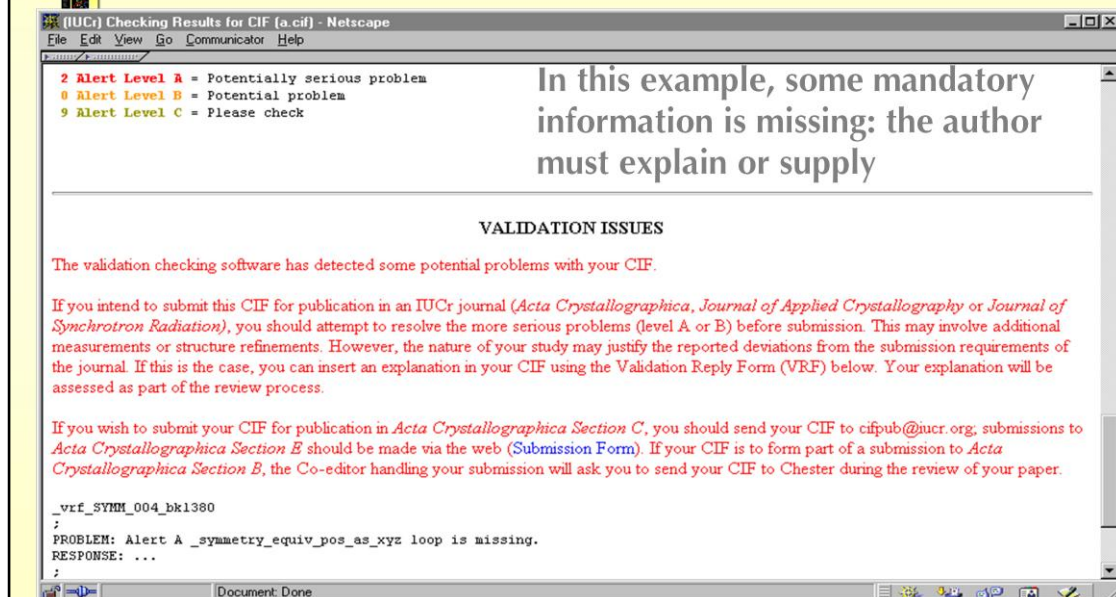
Minor problems/errors

In this example, a query is raised about the assigned chirality



In this example the Flack parameter has a high value, suggesting that the structure has been assigned an incorrect chirality (left- or right-handedness; the sense of orientation of asymmetric molecules may not be unambiguously proven by the crystallographic experiment). The attention of the author (and referee) is drawn to this fact, and the author would normally be required to provide a structural model for the inverse enantiomer. The rear window contains a terse report of the problem. A more detailed description of the problem (often with helpful advice on how to solve it) is available through a hyperlink (front window)

Severe problem



Where essential information is missing or a gross error appears to be present, the author is informed in no uncertain terms that the article as it stands is very unlikely to be acceptable for publication. A form is provided in which explanation or justification can be supplied by the author. This form (which is a text file) should be completed by the author and pasted into the CIF that is eventually submitted for publication. Each journal has its own policy for handling such severe warnings. For *Acta Crystallographica Section E*, the structure *will not be published* **unless** the author provides justification against each such severe notice.

The screenshot displays a web browser window with two main panels. The left panel shows the 'Structure Reports Online' website for the International Union of Pure and Applied Chemistry (IUCr). The page title is 'Structure Reports Online'. Below the title, there are navigation links: 'html', 'pdf', 'abstract', 'cif', '3d view', 'structure factors', 'checkCIF', and 'buy'. The 'cif' link is highlighted with a red circle. The main content area shows the title of a paper: 'Bis(μ -pyridinyl-1-oxide)-1 κ O¹:2 κ C²;1 κ C²:2 κ O', the authors 'R. Fandos and M. D. Walter', and the publication details 'Acta Cryst. (2006). E62, m264-m266 [doi:10.1107/S1600502906000000]'. The right panel shows a 3D molecular model of the compound, rendered in a ball-and-stick representation. The model is displayed within a unit cell, with axes labeled 'a', 'b', and 'c'. The atoms are colored by element: carbon (grey), oxygen (red), nitrogen (blue), and phosphorus (pink). The model is viewed from a perspective that shows the three-dimensional arrangement of the atoms. The browser window is titled 'Mozilla Firefox' and shows the URL 'http://journals.iucr.org/structure/2006/e62/m264-m266.html'.

Although the editors and referees invest much effort in the peer review process, the availability of a full set of accompanying data provides added value for the reader of the article. For example, for any article published in *Acta Crystallographica Section E: Structure Reports Online*, the reader may assess fully the scientific argument by: (1) reading the text of the article; (2) accessing the full CIF (which will include unpublished data such as the three-dimensional atomic coordinates, and a complete listing of bond lengths and angles); (3) reviewing key indicators and the validation report (in this example the author's response to a significant problem identified in the validation review is presented); (4) retrieving the primary experimental data to allow an independent redetermination of the structure; or (5) visualizing and manipulating the data in a crystallographic application of choice. The example shown is a program distributed by the Cambridge Crystallographic Data Centre allowing full visualisation of the six-dimensional model (displacement ellipsoids may be visualised as well as the three-dimensional positional coordinates), generation of the crystal lattice, exploration of molecular geometry and intermolecular bonding *etc.*

Extending the approach

- Consensus in small-molecule crystallographic community
- Emerging standards in macromolecular crystallography
- *actabiostandards*

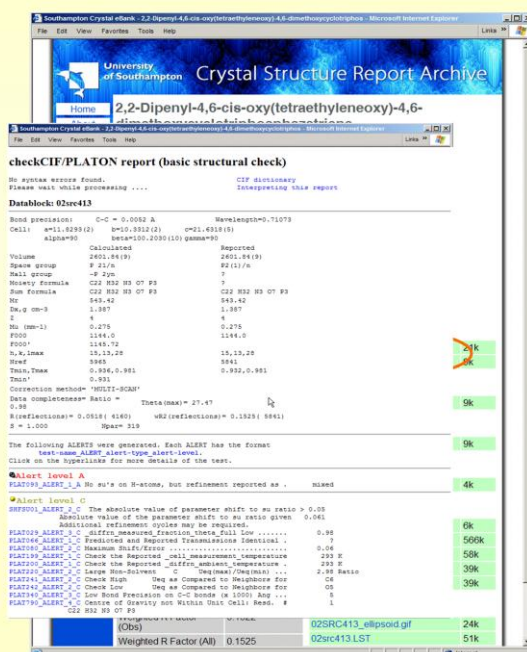


Crystallography
Journals
Online


The high level of automated handling and semi -automated evaluation of the quality of the data has gained consensus approval in the small -unit-cell (small-molecule and inorganic) crystal structure community. The IUCr journals are now working with the macromolecular structure community to bring similar standards to the field of protein and nucleic-acid structures. The scientific challenges are substantial: the experiments are inherently more complex, the quantity of data much greater, and the derived structures more complicated and difficult to characterise. Nevertheless, the availability of a well-defined data description provides the necessary reference point against which to set well-characterised benchmarks. The *actabiostandards* discussion list managed by the editorial boards of *Acta Crystallographica Sections D* and *F* is working towards the development of suitable data quality and description standards. (As an aside, the animated ‘covers’ of the electronic -only IUCr journals *Acta Cryst. Sections E* and *F* are generated directly from submitted CIF data – they are not pre-generated graphics supplied by the authors.)

Extending the scholarly publication paradigm

- ePrints repository
- OAI-PMH
- Standard metadata
- All data
- Links to publication
- Rights
- Quality



The volume of new crystal structure determinations is now so high that not all structures can be reported in the literature. An additional benefit of a standard such as CIF is that separate institutional or laboratory repositories are beginning to be established that store and disseminate crystallographic information in the same format as used by the journals. Among the most active of these laboratory repositories is eBank, the University of Southampton/National Crystallography Service server. Here, subject to appropriate permissions, details of structures determined within the laboratory are made available under an open-access architecture. The platform supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), allowing automated harvesting of the metadata describing individual entries. Potential developments from this repository include: ingest of entries to the Cambridge Structural Database and other curated structural databases; federated searching and archiving services; linking to published articles. The IUCr is helping to define suitable metadata to allow automated processing by chemical or other scientifically-aware information services, drawing on the vocabularies provided by the CIF dictionaries. The eBank implementation provides links to *all* the supporting data collected during the experiment (including links to the raw data images, archived in a national large-scale data facility). IUCr journals plan to provide links between subsequently published articles and the corresponding eBank records. Use of open protocols and data exchange standards results in few technical barriers to the development of this as a new publication medium, complementing traditional publications. However, rights to ownership and dissemination of the data (collected and analysed in a client/provider relationship) need careful consideration and handling. In a framework where the published information has not necessarily undergone peer review, the eBank developers have addressed quality concerns by analysing each structure with the IUCr *checkCIF* software and providing a link to the relevant report.




What is innovative?

- Rich data description standard
- Interoperability with publication markup
- Automated support for applying data quality policy
- Community involvement in quality issues
- Reader: interaction with the scientific argument

Crystallography
Journals
Online

In summary, what is especially innovative about the IUCr's development and application of CIF and *checkCIF*? It has permitted an extraordinarily close synergy between the data contributing to a scientific model, and the discussion of that model. It has greatly improved the efficiency of the data/publication cycle, with immense benefits in conducting the research itself, in writing up the results and submitting them for publication, in peer review and quality assessment, and in the community recognition of important criteria for determining quality. Perhaps most importantly, it has allowed the reader – the end user of the scientific publication – truly to interact with the published model: not only by passive reading, but through active analysis, interrogation, reanalysis and reuse of the associated data.



Future prospects

- Increased automation ('next generation dictionary definition language')
- Engagement of other communities (macromolecular structures, powder diffraction, incommensurate structures, NMR, chemical properties...)
- More integration between data and literature
- Interoperability with domains beyond crystallography

Crystallography
Journals
Online

In the crystallographic domain, work is actively under way to extend the functionality of the CIF dictionary definition language and so to improve the efficiency and power of machine processing of the contents of CIFs. The work we have done on small-unit-cell structures will be extended to other areas of structural science. And, although crystal structures are a reasonably homogeneous class of scientific object that lend themselves to an unusually high level of uniformity and efficiency in their handling, we hope that we can encourage other domains to follow our lead and begin to develop similar systems for handling other types of data -centric scientific descriptions. Already in the journals where CIFs are used as supplementary documents rather than as the publication medium itself, we have demonstrated how such data files can be more closely integrated into the experience of using online scientific journals. We look forward to the day when we can begin to link directly from crystal structure reports to related scientific publications in other areas of the electronic scientific knowledge society.