

# *pubCIF*: software for editing, validating and formatting crystallographic information files

Simon P. Westrip

Technical Services, The Walled Garden, Iscoyd Park, Iscoyd, Shropshire SY13 3AR, England. Correspondence e-mail: sw@pubcif.co.uk

*pubCIF* is an application designed for creating, editing and validating crystallographic information files (CIFs) that are used in journal publication. It validates syntax and dictionary-defined data attributes through internal routines, and also provides a web interface to the *checkCIF* service of the International Union of Crystallography (IUCr), which provides a full crystallographic analysis of the structural data. The graphical interface allows users to edit the CIF either in its 'raw' ASCII form (using a text editor with context-sensitive data validation and input facilities) or as a formatted representation of a structure report (using a word-processing environment), as well as *via* a number of convenience tools (*e.g.* spreadsheet representations of looped data). Beyond file and data validation, *pubCIF* provides access to resources to facilitate preparation of a structure report (*e.g.* databases of author details, experimental data, standard references *etc.*, either distributed with the program or collected during its use), along with tools for reference parsing, spell checking, structure visualization and image management. *pubCIF* was commissioned by the IUCr, both as free software for authors and as a tool for in-house journal production; the tool for authors is described here. Binary distributions for Linux, MacOS and Windows operating systems are available.

© 2010 International Union of Crystallography  
Printed in Singapore – all rights reserved

## 1. Introduction

An essential element in the design of the crystallographic information file (CIF; Hall *et al.*, 1991) was an ability to convey the full text of a structure report for publication in the research literature. Since its adoption by the International Union of Crystallography (IUCr) as an information exchange standard, CIF has been used as a vehicle for submitting such articles, in whole or in part, to a number of IUCr journals. Since CIF uses the ASCII character set, it has always been possible to edit CIFs with simple text editors, but authors have needed to understand how CIF syntax delimits a data value from a data name, and to learn the encoding used for special characters and symbols, subscript, superscript, italic or bold markup. There has always been the risk of corrupting or destroying the integrity of the file contents through inadvertent syntax errors.

*pubCIF* has been developed as a desktop application to help authors to create or edit CIFs for publication, whether they are familiar with CIF or new to the format. It provides an interactive editing environment, but goes far beyond this to support file and data validation, CIF dictionary look-up and checking, reference management and processing, and image management and visualization. Its user interface has been designed for both the CIF 'expert' and the 'novice', providing a dual editing interface (a 'raw' ASCII CIF editor and a word-processing environment). Context-sensitive menus provide access to CIF dictionary data, while publication 'wizards' facilitate data input, calling upon internal and external resources where available. CIF syntax and data validation are employed throughout, much of it 'as you type'.

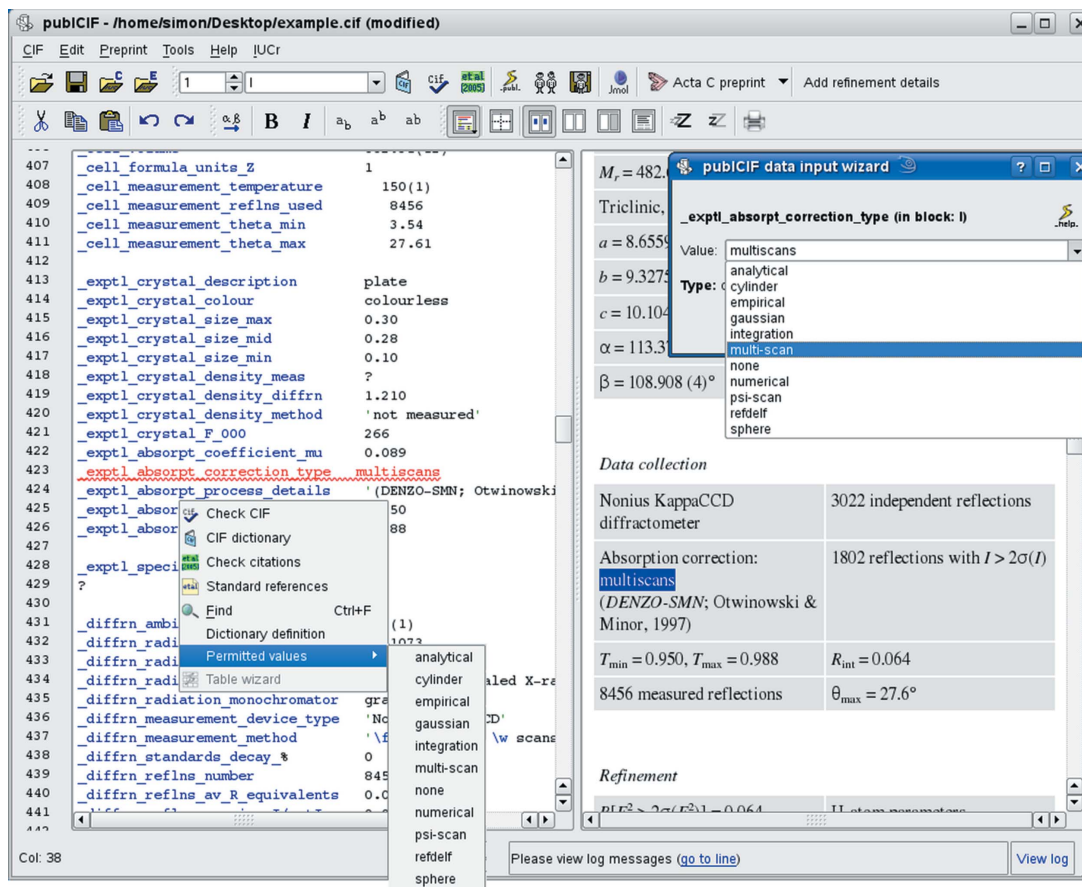
The public version of *pubCIF* described herein is optimized for use with single-crystal and powder CIFs that describe small-molecule

or inorganic structures and is tailored to the publication requirements of *Acta Crystallographica Sections C* and *E*.

## 2. User interface

Fig. 1 demonstrates some aspects of the editing environment. By default, two windows show complementary views of the file. One is the unmodified ASCII content of the CIF (the 'raw' CIF in the 'CIF window'); the other (the 'preprint window') shows a formatted version in the style of one of a number of user-selectable journal formats, or in a generic rendition that displays all the items that might form part of a published article (according to IUCr submission requirements). The user may change the relative position of these windows on the screen (placing the preprint window to the left of, or above or below the CIF window). It is also possible to hide the CIF window entirely, so that an author unfamiliar with CIF can work entirely in a word-processing mode.

In general, the text cursors in the two windows are synchronized, and the user may edit text in either window; the complementary view is updated automatically 'as you type'. This synchronization is especially useful for locating data items within the CIF, which may contain far more data than are presented in the preprint window; moreover, the order of data items within a CIF is largely arbitrary, so the ordered rendition in the preprint window facilitates rapid location of the CIF data source on a single click of the mouse. This synchronization can be disabled to allow free navigation of one window while maintaining the cursor position in the other (*e.g.* in order to type in a text section of the preprint window while viewing experimental data in the CIF window). Furthermore, it is possible to decouple the two views completely to work solely in the CIF window; this is recom-



**Figure 1**

General view of the synchronized *pubCIF* plain-ASCII ('raw') CIF and formatted ('preprint') editing windows, along with a context-sensitive mouse menu and a data-input pop-up. In this example, the status bar at the bottom of the window has alerted the user that there is a problem with an item: the value of `_exptl_absorpt_correction_type` is not one of the dictionary-defined permitted values. The data-input wizard (on the right, called by double-clicking the item in the preprint window) provides a list of the permitted values, as well as links to the dictionary definition if required. The same information is accessible *via* the mouse-activated menu when working in the CIF window. The offending item is highlighted in the CIF window using a wavy underline.

mended when performing complex edits of the CIF (for example, involving rearrangement of data items).

In the following sections, the preprint window will be described first (§2.1). The role of the CIF window will then be explained (§2.2), before a description of the content-validation functionality and associated tools (§3).

### 2.1. The preprint window

The preprint window presents the user with a formatted document in a basic word-processing environment. The main text sections of the paper can be written and edited directly in the preprint window. A formatting toolbar allows the insertion of markup for bold and italic styles, and subscripts and superscripts. Special characters and symbols are listed in a pop-up window accessible from the formatting toolbar.

Tables of experimental data are synthesized from individual data items in the CIF and appear in the preprint window on a tinted background. Because the formatted rendition of the data item is not necessarily exactly the same as the associated CIF source (*e.g.* many numeric values are rounded for publication purposes, CIF codes are translated to meaningful phrases *etc.*), the user may not edit such tinted areas directly. However, a single click on the item will highlight its value in the CIF window, while double-clicking will activate an editing widget in which the value of the data item can be modified without disturbing any auto-generated text or formatting associated

with it. If the allowed value of a data item is constrained by its CIF dictionary definition, the editing widget restrains the user from entering invalid content. Fig. 1 shows a pop-up editing widget that allows the user to select a valid code for the CIF data field `_exptl_absorpt_correction_type`. If the selected data item is normally found looped with other related items, the editing widget takes the form of a spreadsheet (Fig. 2), where all the related data can be edited in a tabular environment. A particular advantage of the spreadsheet widget is that it allows the user to add entire new rows or columns (the latter operation corresponding to adding a new data item to the CIF loop), and will provide drop-down lists of permitted values if appropriate (either read from the dictionary or taken from related data items, *e.g.* atom-site labels).

In addition to these standard editing widgets, a number of data-input 'wizards' are also available (*e.g.* double-clicking the author section will activate a wizard for inputting and editing the author details; see §3).

### 2.2. The CIF window

The CIF window displays the CIF as a raw ASCII file. It was originally designed as a basic plain-text editor, *i.e.* with no syntax highlighting or text formatting. However, in response to user requests and recognizing the merits of any mechanism that facilitates data recognition, especially as CIFs are likely to become more complex in

the future (with extended data types and dictionaries that contain interpretable code to derive data values), a syntax-highlighting mode has been implemented recently (version 1.9.6). Syntax highlighting is

both optional and customizable (the user can select the colours employed and whether to render marked-up text in an appropriate font, *i.e.* bold, italic, superscript, subscript).

The CIF window provides access to item-specific information *via* a context-sensitive right-mouse-button menu (Fig. 1). This provides access to dictionary data for the item, including a list of permitted values if appropriate, which can be inserted directly from the menu.

Whenever the CIF window is clicked or edited, *publCIF* attempts to identify the relevant data item and validate it against a CIF dictionary, as well as check the syntax of the data structure. Any problems are reported in a status bar at the bottom of the main interface. In addition, if the item is also represented in the preprint, the preprint window will be updated.

### 3. Content validation

An important function of *publCIF* is to enable a user to create a file that is 'valid' against a number of criteria.

#### 3.1. Syntax checking

Whenever a CIF is opened or closed using *publCIF*, the syntax of the CIF is checked using the external program *vcif* (McMahon, 2005). Problems are reported in a pop-up log window, and links are provided from this log window directly to the line or lines that triggered the reported syntax errors.

During use of *publCIF*, whenever a CIF item is accessed, either programmatically to render it in the preprint or by user interaction with that item, the syntax of the data item is checked and any errors or warnings are printed in the status bar at the bottom of the interface as well as in the pop-up log window (accessible from the status bar).

A full syntax check can be called at any time using a toolbar button, which will rerun *vcif* as well as *publCIF*'s checking routines.

#### 3.2. Dictionary checking

*publCIF* will scan a CIF for references to conformant dictionaries in the `_audit_conform_dict_` loop, and attempt to load any such dictionaries. Otherwise, it will load the core CIF dictionary. If the user is working in the CIF window, the status bar will report warnings or errors if an unrecognized data name is entered, or if the type or value of data entered for a specific data item contradicts the corresponding dictionary definition. If the user is working in the preprint window, changes to non-text

	Donor (D)	H atom	Acceptor (A)	D...H distance	H...A distance	D...A distance	D...H...A angle	Symmetry A	Publish
1	_geom_h...	_geom_h...	_geom_h...	_geom_h...	_geom_h...	_geom_h...	_geom_h...	_geom_h...	_geom_h...
2	C2	H2B	O21	0.99	2.45	3.39(2)	159	1_455	n
3	N1	H1B	O33	0.92	1.68	2.60(4)	173	1_445	n
4	N1	H1A	O23	0.92	1.85	2.760(11)	172	1_445	y
5	N1	H2A	N11	0.92	2.05	2.799(8)	138	.	y
6	N11	H2B	O34	0.92	2.18	3.06(2)	158	1_445	n
7	N11	H3A	O24	0.92	2.21	3.094(8)	161	1_445	y
8	N4	H51A	O32	0.92	2.29	3.17(4)	159	.	n
9	N4	H51B	O22	0.92	2.17	3.046(15)	160	.	y
10	N8	H51C	O31	0.92	1.87	2.78(5)	171	.	n
11	N8	H52A	O21	0.92	1.78	2.702(14)	179	.	y

Notes on hydrogen-bond geometry (\_lucr\_geom\_hbonds\_special\_details)

N1...H1B...O33(x-1, y-1, z)

(a)

	1	2	3
1	(II)	.	.
2	(N11-C16)^(N11^+-C16^+)	15.8 (2)	.
3	.	.	.
4	(III)	.	.
5	C13<D>B</D>-C14<D>B</D>-C17<D>B</D>-C27<D>B</D>	-1.7 (6)	C13<D>F</D>-C14<D>F</D>-C17<D>F</D>-C27<D>F</D>
6	C14<D>B</D>-C17<D>B</D>-C27<D>B</D>-C24<D>B</D>	-178.4 (4)	C14<D>F</D>-C17<D>F</D>-C27<D>F</D>-C24<D>F</D>
7	C17<D>B</D>-C27<D>B</D>-C24<D>B</D>-C23<D>B</D>	102.9 (5)	C17<D>F</D>-C27<D>F</D>-C24<D>F</D>-C23<D>F</D>
8	.	.	.
9	C13<D>D</D>-C14<D>D</D>-C17<D>D</D>-C27<D>D</D>	4.5 (4)	C13<D>H</D>-C14<D>H</D>-C17<D>H</D>-C27<D>H</D>
10	C14<D>D</D>-C17<D>D</D>-C27<D>D</D>-C24<D>D</D>	179.5 (3)	C14<D>H</D>-C17<D>H</D>-C27<D>H</D>-C24<D>H</D>
11	C17<D>D</D>-C27<D>D</D>-C24<D>D</D>-C23<D>D</D>	-1.4 (4)	C17<D>H</D>-C27<D>H</D>-C24<D>H</D>-C23<D>H</D>
12	.	.	.

Table footnotes (if necessary)

Symmetry code: (i) y, x, -z.

(b)

**Figure 2**

(a) A spreadsheet widget allowing the user to edit a CIF loop. In this example, the user is changing the H atom of a hydrogen bond. Entries flagged for publication show up in the editing widget on a coloured background. Hydrogen-bond data should always be archived using the dictionary-defined data items, even if the data are not to appear in the published report. The widget recognizes these items and provides drop-down lists of permitted values (as read from the dictionary or taken from related data in the CIF, *e.g.* atom labels, symmetry codes). (b) A similar widget, but allowing non-standard tables to be prepared and stored in the CIF. In this case the data items are arbitrary inasmuch as their definition provides no information about the data value (thus excluding the use of the data for anything but publication purposes). In the example shown, the user is preparing a hydrogen-bond comparison table. The data shown have been imported by copying a table of hydrogen bonds from a web page describing a related structure and then clicking the 'Import data' button; *publCIF* has converted the copied data to CIF format and populated the table cells. The next step might be to copy tabulated data from the current preprint and import it, then use the widget's table manipulation tools to rearrange the rows/columns/cells. This widget provides fields for adding a table heading, headnotes and footnotes.

data fields are normally managed through pop-up data-entry wizards which enforce the constraints recorded in the dictionary. Sometimes this can be done in an automated way. For example, consider a CIF that initially contains a table of bond lengths with values of 'Yes' for the `_geom_bond_publ_flag`. This tag is case-sensitive, and so validating the file will raise several errors. However, if the user opens the loop in a spreadsheet editing widget (simply by double-clicking on an entry in the loop), all the 'Yes' values will automatically be normalized to the permitted 'yes' as the spreadsheet is loaded. The result can be immediately saved to achieve what could otherwise involve a tiresome editing procedure.<sup>1</sup>

In the CIF window, the user may also access a context-sensitive menu by right-clicking the mouse over a CIF data name. This menu allows the user to read the dictionary definition of the currently selected item, or to open up a dictionary browser window in which all other items can also be consulted (Fig. 3). The dictionary browser has internal hyperlinks allowing easy navigation of related dictionary items, and it can also be used to drag-and-drop new data items into the CIF editing window.

The 'as-you-type' dictionary validation functions can also be run on demand using the same toolbar button as described in the preceding subsection.

### 3.3. Data validation by *checkCIF*

If an internet connection is available, the user can access the IUCr online *checkCIF* service (<http://checkcif.iucr.org>) through a simple menu [see Strickland *et al.* (2005) for an account of *checkCIF*]. The current version of the CIF is uploaded directly to the *checkCIF* server, and the result (including an atomic displacement ellipsoid plot) is displayed in a pop-up window. If the *checkCIF* report contains 'type A' alerts (indicating significant outliers from expected chemical behaviour), a validation response form will be included in the report. *publCIF* offers a simple button in the report window, 'Add VRF to CIF', which pastes the validation response form into the CIF currently being edited. This allows the user very easily to provide a textual commentary describing the reason for the outliers identified by *checkCIF*. Inclusion of such a validation reply form is mandatory for submission of articles containing 'type A' alerts to *Acta Crystallographica Sections C and E*, but the ease of including it through the *publCIF* interface to *checkCIF* may encourage authors to provide a helpful annotation even where it is not obligatory.

### 3.4. Publication-oriented validation

As well as the extensive CIF syntax, dictionary compliance and crystallographic data validation described above, a number of procedures allow the dynamic checking of content with respect to general preparation of a scientific report. As already alluded to above, *publCIF* provides a number of 'wizards' for data entry and management, each employing validation routines appropriate to their purpose. The aforementioned author-details wizard is one example, which helps to maintain consistency between submissions from the same group of authors.

One of the most useful publication-oriented tools allows the user to scan the textual material and check that citations in the text match entries in the reference list. This, and the author-details wizard, will be described in more detail in the following section.

<sup>1</sup> In the latest version of *publCIF* (version 1.9.6), such auto-correction (*i.e.* correcting data where the only dictionary violation is the text case used for a value) can be performed when the CIF is opened.

## 4. Resource management

*publCIF* works effectively as a standalone CIF editor, but its particular strengths in editing files for publication include storage and management of external information associated with one or more published articles, and the collection of data from an open CIF for use with other CIFs. For example, the data-item input widgets will collect textual data (*e.g.* diffractometer names) and make such data available for future use. Further examples are given below.

### 4.1. Author details

A small database is maintained that contains the names, initials, addresses and explanatory footnotes of authors (*i.e.* individuals listed in the `publ_author_` fields). The use of this database allows author names to be entered consistently and with proper markup across a number of CIFs. If the user double-clicks on the author details in the preprint window, a structured interface to the author database is presented. This interface is also invoked if the user launches the 'paper creation wizard' utility to create a new publication from a data-only CIF or from an empty file.

The database is populated by using a toolbar button to save the current authors. This same button provides an administrative interface to the database.

If an internet connection is available, the author-details wizard also provides a widget to search the World Directory of Crystallographers by author surname, and include the returned results in the current author list if required (Fig. 4).

### 4.2. Citations

A common source of inconsistency in publications is the reference list. *publCIF* provides a tool to parse the reference list, attempting to identify the authors and journal name, volume and page information, and then scan the discursive text in the CIF, attempting to find citations and match them against entries in the reference list. When an internet connection is available, the reference-parsing process will



**Figure 3**  
The CIF dictionary browser window implemented in *publCIF*. The data items are shown in a 'tree view' on the left; a formatted representation of the dictionary description of the selected item is shown on the right.

attempt to validate references to articles in IUCr journals against the Crystallography Journals Online database (using an automated version of the search facilities available at <http://journals.iucr.org/services/search.html>). A summary report is presented upon completion of these checks and any possible ambiguities are highlighted in the preprint window. More details are written to the log window.

A comprehensive list of standard references is also available for searching and pasting into the CIF.

Citation management is an area of ongoing development, and is likely to benefit from wider use of external resources. Indeed, enhancements have already been made to the version of *pubCIF* that drives some of the IUCr's online author services (see §6).

### 4.3. Graphical images

Most CIF-based articles will include figures, chemical schemes or other illustrations. *pubCIF* allows an author to preview graphics files associated with the CIF in the preprint window, and will keep track of associated graphics files between editing sessions.

In addition, structure visualization using *Jmol* (<http://www.jmol.org>) is provided by implementing a 'local' version of the IUCr's online enhanced figure toolkit (Strickland & McMahon, 2008; McMahon & Hanson, 2008). When *Jmol* is launched *via* the toolbar button or *via* one of the image menu items, the user is presented with an interactive visualization of the structure in their default internet browser. The interface will 'look and feel' like a web service, but no internet connection will be necessary: *pubCIF* acts as the 'web server', managing the data transfer between the open CIF and the toolkit 'web pages' in the browser. This feature not only provides a convenient means to inspect the structure graphically, but can be used to create static images for inclusion in the preprint. Moreover, any *Jmol* scripts used to create an image or three-dimensional model can be stored in the CIF so that the model is automatically generated when the CIF is opened again in *pubCIF* or any tool that is able to read the scripts (e.g. the IUCr's online *printCIF* service).

Such integration with other useful crystallographic software is another area of ongoing development of *pubCIF* (see §6).

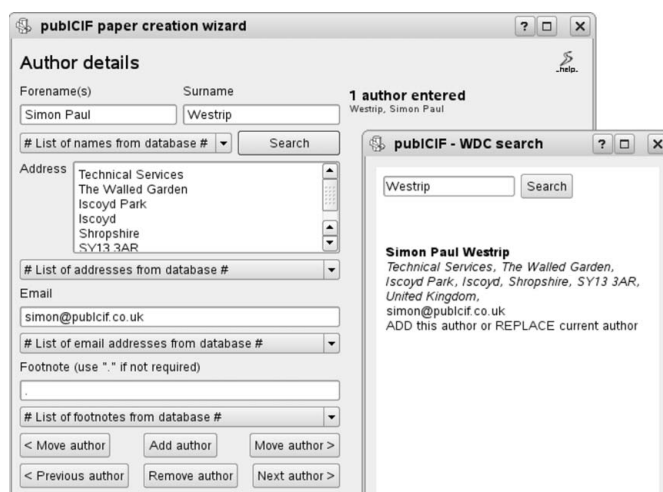
### 5. Documentation and availability

Extensive user documentation is provided within the application, presented *via* context-sensitive pop-ups and mouse menus, and in a manual that can be browsed in a separate window.

*pubCIF* is written in C++ using the *Qt* libraries (<http://qt.nokia.com/>). Binary distributions for Linux, MacOS and Windows operating systems are available from <http://pubcif.iucr.org>. *vcif* (McMahon, 2005), *Aspell* (for Windows; <http://aspell.net>) and *Jmol* (<http://www.jmol.org>) are currently bundled with the distribution. *pubCIF* and all libraries distributed with it are free of charge for non-commercial use, under the terms of the copyright notices and licences included in the packaging of the software.

### 6. Ongoing development

*pubCIF* is used in-house by the IUCr in its publishing activities to edit CIF-based articles and generate SGML, HTML and PDF output as part of the journal submission and production process. As such, it is subject to ongoing development, both to reflect any changes to publication requirements and to improve the tool in general. When any of these developments are thought to be useful for authors they will be implemented in the public version.



**Figure 4** The author-details wizard, like other data-input tools in *pubCIF*, provides access to data collected from other CIFs (in this case author names and addresses), as well as online resources (in this case the World Directory of Crystallographers can be searched and the results imported directly into the CIF).

Equally as important as issues arising from in-house use is feedback from authors, especially as editorial use tends not to encompass many of the features that are designed for authors. Bug reporting is essential; please send suggestions and feature requests to [support@iucr.org](mailto:support@iucr.org).

Although the public version of *pubCIF* is currently optimized for use with single-crystal and powder CIFs that describe small-molecule or inorganic structures and is tailored to the publication requirements of *Acta Crystallographica Sections C* and *E* (especially with regard to encapsulating the entire structure report in a single CIF), the program is able to generate output for other publication purposes and not only for the aforementioned disciplines. For example, tables in rich-text format (RTF, recognized by most word processors) can be prepared from macromolecular CIFs and modulated-structure CIFs, as well as from the 'core' single-crystal and powder CIFs. This functionality can be seen in some of the recent enhancements to the IUCr's online CIF-based author services, where *pubCIF* serves to process the uploaded CIF and generate both the HTML interface and the downloadable output (RTF and PDF).

It is envisaged that this more flexible (compared with *Acta Crystallographica Sections C* and *E* requirements) use of CIF as a data source for scientific literature will influence future developments of *pubCIF*, especially as CIF becomes more powerful with the introduction of methods-driven dictionaries, yet, at the same time, inevitably less accessible in its raw form.

Beyond publication-oriented functionality, increased implementation of internal crystallographic analysis would provide added value, so that *pubCIF* may find an educational role or be useful as a convenient 'generic' means of examining any structure for which a CIF is available. Further integration with other software will be pursued where practicable and beneficial.

This project was supported by the IUCr and managed by Peter Strickland (Managing Editor), to whom the author expresses his sincere gratitude. Suggestions from a number of users, including the referees, have resulted in improvements of the program. In particular,

---

Sarah Froggatt (Technical Editor, IUCr Journals) is thanked for her diligent testing and invaluable feedback throughout the development of *publCIF*.

### References

- Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* **A47**, 655–685.
- McMahon, B. (2005). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, pp. 499–525. Dordrecht: Springer.
- McMahon, B. & Hanson, R. M. (2008). *J. Appl. Cryst.* **41**, 811–814.
- Strickland, P. R., Hoyland, M. A. & McMahon, B. (2005). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, pp. 557–569. Dordrecht: Springer.
- Strickland, P. R. & McMahon, B. (2008). *Acta Cryst.* **A64**, 38–51.