AN INTEGRATED WEB RESOURCE FOR CRYSTALLOGRAPHY

Brian McMahon

International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK
Email: bm@iucr.org

ABSTRACT

A recurring theme during the CODATA 2000 conference (Lake Maggiore, Italy, 15 - 19 October 2000) was the increasing convergence in data-rich branches of science between the storage and retrieval of data and the publication of conclusions drawn from the data. Web publishing technologies facilitate access to publications and data through the same interfaces and tools. For crystallography, the ability to deliver the experimental data alongside the research commentary offers tremendous advantages. A structured file format has been developed that allows not only submission of a research article accompanied by a complete supporting data set, but also automated validation of the description of the crystal structure reported in the article against the accompanying data. Such validation is an important component of the review process, and encourages better-quality publications. The adopted format is different from XML, but shares some of the properties of that markup language; and suggests the improvements in quality that might result in other subject areas from the adoption of similar methodology. The International Union of Crystallography fully exploits the convergence of publishing and data-handling technologies in its online journals and associated Web site.

Keywords: Data exchange standards, crystallography, World Wide Web, online journals.

1 INTRODUCTION

Crystallography is the branch of science devoted to the study of molecular and crystalline structure, with far-reaching applications in chemistry, physics, mathematics, biology and materials science. The International Union of Crystallography (IUCr) was founded shortly after the Second World War to represent the interests of crystallographers and was incorporated into the International Council of Scientific Unions (ICSU: now the International Council for Science) in 1947. Its major aims, by statute, are:

- To promote international cooperation in crystallography
- To contribute to the advancement of crystallography in all its aspects, including related topics concerning the non-crystalline states
- To facilitate international standardisation of methods, of units, of nomenclature and of symbols used in crystallography
- To form a focus for the relationship of crystallography to other sciences

It will be apparent that its charter has a broad interdisciplinary appeal, and that there is a strong incentive to develop methods of intercommunicating technically and at a personal level with other disciplines. When the IUCr was founded, it undertook the publication of its own journal, *Acta Crystallographica*, to ensure that the cross-disciplinary nature of the field was properly represented. From the beginning *Acta* carried, in addition to its scholarly research papers, notices of conferences, publications, commercial products, book reviews, obituaries, and reports on the work of the commissions and committees, scientific and administrative, of the Union itself.

The IUCr has always taken very seriously the needs of its community for education, publication of research results and organisation of the data that are core to the discipline. More than half a century later the IUCr remains true to its original goals. Traditional publishing activities embrace seven journal titles,[1] a major series of standard reference volumes, several book series, an informal Newsletter and miscellaneous other publications. The ubiquity of the Internet and its many technologies for distributed authorship and distribution of content now make it a natural choice as an

[1] The full list of journal titles published in 2002, together with first year of publication, is: *Acta Crystallographica Section A: Foundations of Crystallography* (1968), *Acta Crystallographica Section B: Structural Science* (1968), *Acta Crystallographica Section C: Crystal Structure Communications* (1983), *Acta Crystallographica Section D: Biological Crystallography* (1993), *Acta Crystallographica Section E: Structure Reports Online* (2001), *Journal of Applied Crystallography* (1968), and *Journal of Synchrotron Radiation* (1994).

information conduit. In recent years the IUCr has established a powerful and comprehensive information resource on the Web that currently includes the full content of all its research journals, and a substantial corpus of other research and informal material.

Among the key design features of the IUCr web site are:

- logical structuring of information content;
- clean navigation within content areas, aided by optional navigation frames;
- an architecture of static files in a deep directory hierarchy optimised for widespread mirroring;
- a central server to generate dynamic content, with tracking of the referring page to allow dynamic generation of links to other (static) locations on the invoking mirror;
- delegation of authorship for specific content areas;
- a hub distribution system allowing automated upload of content from authors and download of content to mirrors;
- efforts to retain persistence of published references to pages by server redirects when content is relocated;
- effective but modest use of graphics to streamline document loading.

It is the purpose of this article to summarise some innovative design features and applications of the IUCr Web site and indicate their value to the crystallographer and structural chemist, and in particular to highlight exciting new developments in the handling of data relevant to the publication. Among the most important elements of the system described here has been the development of a standard information interchange mechanism that permeates the work of the practical crystallographer from experiment through publication and subsequent retrieval and analysis of the published results. This mechanism, the Crystallographic Information File (CIF), comprises a simple but extensible data model and a rich set of data definitions that together form a machine-readable ontology that could act as a model for other scientific disciplines.

2 DESIGN ELEMENTS

The overall design of the site is based on a number of principles. The logical structure is hierarchical, so that topics and subtopics map well onto a filesystem directory hierarchy. In consequence, the root of the document tree is well defined and simplifies HTML links between documents. Common icons or other graphical elements may be stored in a library directory at a well-defined location accessible to all pages. This allows modularity of design, and clusters of pages may be moved up, down or sideways within the document tree as their relative importance changes. A corollary of this design is that authorship and maintenance of particular components of the system can be delegated to external collaborators.

Because the organisation of topics maps so well onto a computer filesystem, most of the content is carried as static files, and therefore mirror servers can be set up with a minimum of effort. Only the ability to serve static HTML and graphics files is required. Therefore it is possible to have mirrors functioning effectively in countries with poor network connectivity and access only to relatively modest computing equipment. Except for the journal articles, most pages deliberately use minimum graphics and no Javascript or Java applets.

2.1 Distributed authorship

Distributed authorship occurs in practice through delegation of responsibility for content to the various committees and commissions which form the working body of the IUCr. Some material is generated by authors with direct access to the main server in Chester, England; but in many cases authors create their HTML pages on a computer local to them, and the pages are fetched by regularly scheduled software jobs using mirroring software such as *wget* and *rsync*.

Authors of the various web pages must conform to a modest set of rules that allow the automated mirroring system to modify automatically URLs that provide links to other parts of the IUCr web site. There are also some requirements to decorate the pages with navigational aids in a particular house style.

2.2 Distributed delivery

Currently the central server at Chester serves pages for redistribution from ten sites: a higher-bandwidth server in the UK, a powerful server in the USA to overcome the historical bottleneck of transatlantic bandwidth; and local sites set up at the request of, and with the assistance of, crystallographic National Committees in Switzerland, France, Sweden, Russia, Japan, Australia, Israel and South Africa. Sites update their contents from Chester on at least a daily basis, using similar software to that involved in the harvesting of material from distributed authors. Each national mirror carries a logo marked with the flag of that nation to identify its location. Except for this one modification, the contents of all the files on all the mirrors are identical.

2.3 Interplay between static content and database queries on central server

While static pages are easy to create, store and redistribute, an effective web site must be able to handle dynamic content, to provide responses to database queries, literature searches and so on. This is achieved by having all requests passed to the central Chester server. When the central server generates HTML pages containing links to other pages within the IUCr web site, the links are sent as relative links, so that a user placing a query *via*, say, the Swiss mirror will be directed to the required pages on the Swiss mirror. To accomplish this, the central server responding to the query must obtain the location of the calling page from the user's browser. Although most common browsers are able to supply this information, in some cases it is unavailable. In that case the central server has the fall-back option of supplying links to the requested pages on the Chester server. An astute user may notice that a result page with a URL such as http://www.iucr.org/iucr-top/a/b/c/d will correspond exactly to the page on the local mirror with URL http://www.*xx*.iucr.org/iucr-top/a/b/c/d where *xx* is the two-letter ISO code for the nation hosting the mirror (*e.g.* ch for Switzerland). While such navigational expertise is not demanded of users, it is another indication of the homogeneity of the site across its mirrors.

3 ONLINE JOURNALS

The online journals of the IUCr strive to attain all the functionality demanded of online publications nowadays. Journal articles have been marked up in SGML (Standard Generalised Markup Language) (ISO, 1986; Goldfarb, 1990) since 1999 or 2000, and for almost all articles published since then a full range of links is available to supporting and related articles, supplementary data sets and other resources. Article headers (containing bibliographic information only) are available for older articles, and page images exist for all articles published in all IUCr journals since 1948.

All journal articles produced from SGML are available also in HTML format. Mathematical equations are embedded as graphics in the flow of the HTML pages, because structured mathematical markup is still not yet fully developed in the web environment. However, the source mathematics is marked up using the standard language for the purpose, TeX (Knuth, 1984). In principle structured mathematical information may be presented on the web at a later date as the appropriate technology matures. Users have the choice of browsing the HTML text or downloading page images in PDF format. Experience suggests that the HTML representation is invaluable for locating articles of interest and for following links to other publications or data repositories; but that users still prefer to print off the traditional typeset representation for in-depth reading away from the computer screen. In the HTML version of an article, a drop-down menu in the page header frame allows the user to jump at once to any section of the article. Internal hyperlinks allow the user to scroll immediately to a referenced table, figure, equation or literature citation. Citation details appear at the cursor without the necessity of actually scrolling to the reference list.

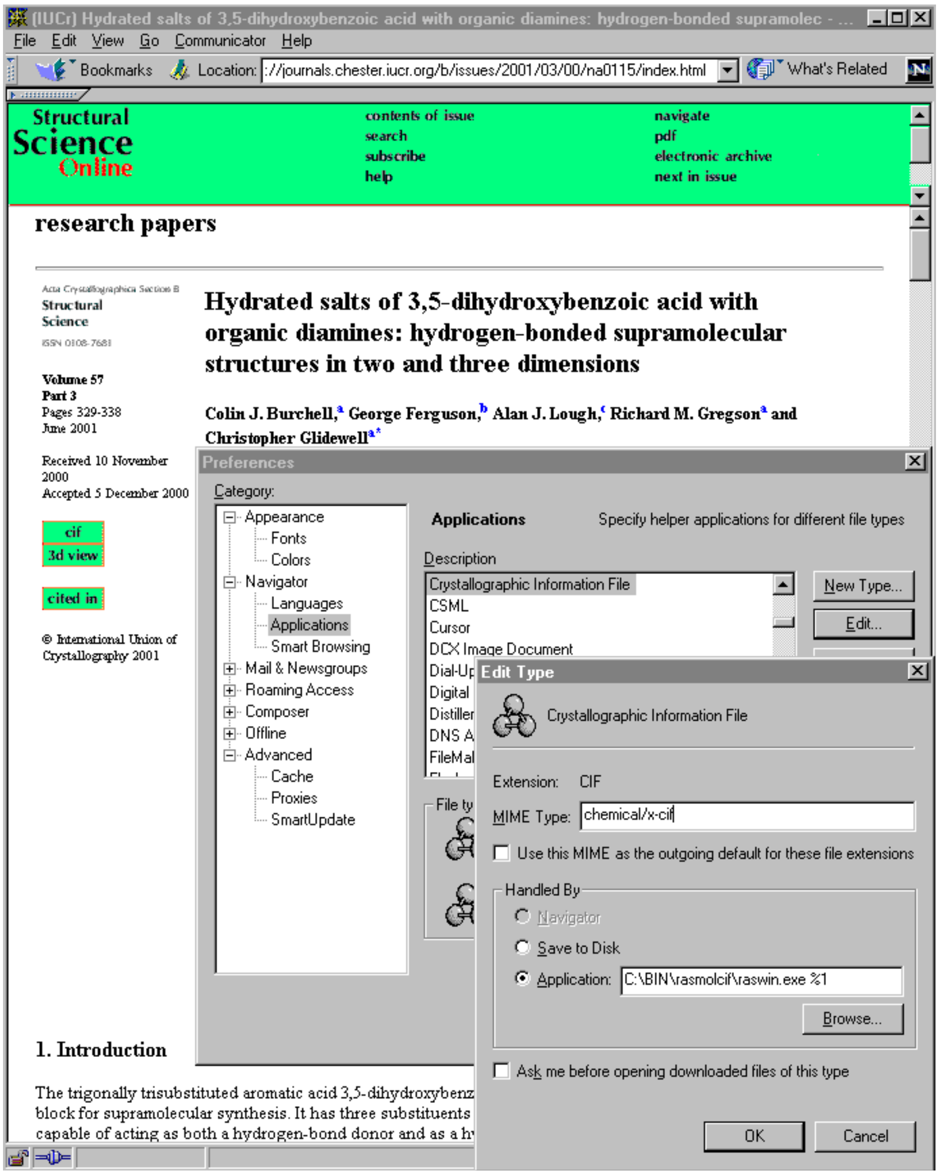
The tables of contents and author indexes are fully searchable to 1948. The entire contents of older issues of the journals since 1948 are accessible online as PDF page images, as the result of a large-scale project in collaboration with a centralised digitisation service for the UK academic community (Higher Education Digitisation Service, n.d.). This digitisation project has involved almost 1000 journal issues, yielding almost 47,000 articles across 175,000 journal pages.

While the full text of any article (including figures and tables) is immediately available to subscribers, many articles have supporting or supplementary materials useful to the interested reader. Journal articles reporting crystal or molecular structures utilise a standard interchange file format developed for the purpose. Crystallographic Information Files (CIF) are generated by ubiquitous experimental software, and may be annotated to form a rich adjunct to an article, or, in the case of some journals, the complete content of the published paper. Structures reported for publication by this mechanism are automatically analysed, and indications obtained of the consistency and quality of the results reported. No structural paper is accepted for publication unless its data content has been validated through web or email-based systems. The CIF data files associated with an article may also be downloaded by readers for visualisation, analysis or modelling.

An essential component of online information is its capacity to link immediately to items of related interest. The long established PubMed service for delivering abstracts and article content in the life sciences has recently been complemented by the CrossRef initiative. This is a central resolving service for bibliographic links to the journals of most major publishers of scientific, technical and medical material. These links, and others directly to IUCr articles, are inserted into the article citation list during the editorial production process. There are also links to bibliographic data and abstracts in the *Chemical Abstracts* database. A novel feature of the IUCr journals is that articles also contain links to relevant protein and nucleic acid structures in the Protein Data Bank (Research Collaboratory for Structural Biology, n.d.), and to small-molecule crystal structures in the Cambridge Structural Database (Cambridge Crystallographic Data Centre, n.d.). Discussions and experiments are under way to extend the linking network to entries in inorganic and other crystal structure databases.

For articles that report crystal structure determinations, crystallographic data in CIF format (as described in the next section) are required. Indeed, for the two sections of *Acta Crystallographica* that specialise in the publication of structure determinations, the entire article must be submitted in CIF format. The information that comprises the body of the article is automatically extracted from the submitted CIF and converted to SGML for journal production. A more complete subset of the included data is also extracted and made available for download as supplementary data in CIF format.

For any article, accompanying supplementary or supporting material is accepted in any suitable standard file format. It is a strength of the http protocol that underpins Web technology that any file may be served to a client program (browser) with an accompanying declaration of type that may be used to trigger a specific behavioural response from the browser. A number of file types specific to chemical information purposes have been proposed (Rzepa, Murray-Rust & Whitaker, 1998). The relevant file type for CIF is ‘chemical/x-cif’, and so a user may configure a Web browser to launch a CIF-specific application such as a molecular graphics visualiser upon download of any such file (Figures 1 and 2).



(IUCr) Hydrated salts of 3,5-dihydroxybenzoic acid with organic diamines: hydrogen-bonded supramolec - ...

Location: ://journals.chester.iucr.org/b/issues/2001/03/00/na0115/index.html

research papers

Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Volume 57
Part 3
Pages 329-338
June 2001

Received 10 November 2000
Accepted 5 December 2000

© International Union of Crystallography 2001

Hydrated salts of 3,5-dihydroxybenzoic acid with organic diamines: hydrogen-bonded supramolecular structures in two and three dimensions

Colin J. Burchell,[a] George Ferguson,[b] Alan J. Lough,[c] Richard M. Gregson[a] and Christopher Glidewell[a*]

1. Introduction

The trigonally trisubstituted aromatic acid 3,5-dihydroxybenz
block for supramolecular synthesis. It has three substituents
capable of acting as both a hydrogen-bond donor and as a h

Applications — Specify helper applications for different file types

Description: Crystallographic Information File

Extension: CIF

MIME Type: chemical/x-cif

□ Use this MIME as the outgoing default for these file extensions

Handled By:
○ Navigator
○ Save to Disk
● Application: C:\BIN\rasmolcif\raswin.exe %1

□ Ask me before opening downloaded files of this type

Figure 1. Screenshot of procedure to associate an application with a file type in a typical Web browser. In this example, the 'Edit/Preferences' menu generates a dialogue window in which the user requests that a molecular visualisation program runs when a file of MIME type 'chemical/x-cif' is downloaded by the browser. When a user selects the '3d view' link from an article such as illustrated, the CIF data associated with the article are delivered with a content type of 'chemical/x-cif', prompting the chosen visualisation program to run (Figure 2). If the user selects the 'cif' link from the same page, the identical file is transmitted to the browser, but with declared content type 'text/plain', so that the data contents will be displayed in the browser window as ASCII text (Figure 3).

4 CRYSTALLOGRAPHIC INFORMATION FILE

The Crystallographic Information File (CIF) was commissioned by the IUCr as an archival and information exchange standard (Hall, Allen & Brown, 1991), and has been implemented in many small-molecule and inorganic structure applications for more than a decade. It drives the publication process for several IUCr journals, and is the preferred input format for entries in the Cambridge Structural Database. A more complex variant, the macromolecular CIF (mmCIF), is in use within the macromolecular structural community, and is used as the internal storage format for entries in the Protein Data Bank (Bourne, Berman, McMahon, Watenpaugh, Westbrook, & Fitzgerald, 1997). The current article concentrates on the role of the small-molecule CIF in the IUCr journals, describing its role in publication and validation.

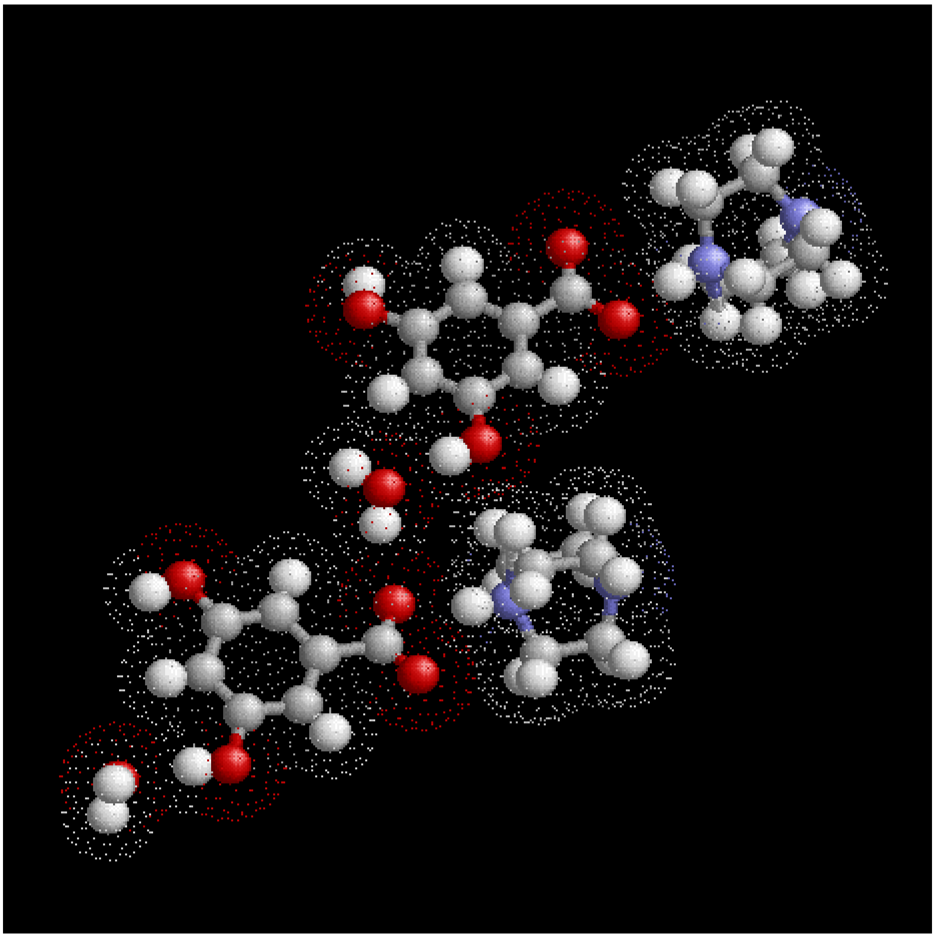


Figure 2. A molecular structure described in a CIF and visualised by a molecular-graphics program (*RasMol*: Bernstein + Sons, 2000).

4.1 File syntax

The syntax of CIF is a proper subset of that specified for the general STAR File exchange mechanism (Hall, 1991; Hall & Spadaccini, 1994). A file is partitioned into one or more *data blocks*, identified by a character string beginning with the five characters **`data_`**. Each data block contains *data items* comprising an identifying *data name* or tag and its associated single or multiple *data values*. Data names are character strings beginning with an underscore character and terminated by white space. Any data name may appear once only within a data block. The associated values may be numbers, words, extended phrases or symbols, or blocks of continuous text, optionally including some typographic markup. Where there is only a single value for an item of data, the value immediately follows its identifying tag, separated by white space but not otherwise constrained to layout. For multiple values, the tag is declared in a header section with a **`loop_`** identifier. Most commonly, related tags are gathered in the same loop header, which acts as a declaration of the columns of a table, and the respective data values follow in sequence.

Figure 3 illustrates these syntactic features with a small extract from a CIF. Note that textual data values that include white space must be delimited by (single or double) quote marks, and that text extending over several lines is delimited by semicolon characters in the first column of the opening and trailing lines.

```
data_I
_chemical_name_systematic
                           'N-(2-hydroxy-1,1-dimethylethyl)benzamide'
_chemical_formula_moiety     'C11 H15 N1 O2'
_chemical_formula_weight     193.24
_symmetry_cell_setting     monoclinic
_symmetry_space_group_name_H-M     'P 1 21/n 1'
```

```
_exptl_special_details
;  The Laue group assignment, the systematic absences and the
   centrosymmetry indicated by the intensity statistics led to
   assignment of the space group uniquely as P2~1~/n (No. 14); since
   refinement proceeded well, it was adopted.
;
loop_
    _geom_bond_atom_site_label_1
    _geom_bond_atom_site_label_2
    _geom_bond_distance
    O1 C7     1.2446(11)
    O2 C9     1.4148(13)
    O2 H1O2   0.890(16)
```

Figure 3. CIF example illustrating various syntax elements.

4.2 Characteristics of the CIF data model

The design of CIF facilitates retrieval of data from a given data block by requesting all values associated with an individual data name or a set of data names. The data names themselves act as pointers to the location of the requested data. No *a priori* knowledge is required of the ordering of data, nor whether a requested data item is actually present in the file. In an arbitrary STAR File, little or no prior knowledge may be required of the relationships between data items; in CIF, however, constraints are applied on how data in tabular form are grouped. In practice, CIF data are grouped by category, and all the data names that are permitted to appear in a single table are classified within the same category. The result is a rather flat data model; hierarchical relationships are poorly defined, but relations between and within categories may be formally defined. The model is similar to that of a relational database, and indeed the data names used in macromolecular CIF (mmCIF) are defined in a strictly relational model.

The definitions of standard data names are stored in external files (also structured as STAR Files) known as *data dictionaries*. A technical committee of the IUCr is responsible for maintaining and commissioning such dictionaries for community-wide use, although there is provision for individuals to add private data names to their own CIFs. The standard dictionaries also include the relationships between data names and the constraints applied to their associated values in terms of data types and permitted ranges or enumerations.

4.3 Relationship between CIF and XML

At a time when XML is increasingly widespread as a vehicle for document markup, it may seem idiosyncratic for the IUCr to promote a domain-specific exchange format. One reason is that CIF is by now very well established within the community, with widespread implementation in commercial and academic software, and substantial user experience. The effort required to educate a generation of crystallographers to work to a common standard demonstrates that such advantages are not easily gained, and their importance is not to be underestimated.

However, a stronger reason is that the existing standard CIF dictionaries have very substantial semantic content. Most extant document type definitions (DTDs) used with XML (or SGML) concentrate on the ordering and hierarchical relationships between document components; there are few examples that fully address the ‘meaning’ of individual components. One noteworthy exception is the Chemical Markup Language DTD (Murray-Rust & Rzepa, 1999).

By contrast, CIF dictionaries define *precisely* the concepts associated with each of several thousand data names, both in a discursive note designed for the human reader, and, where applicable, by machine-readable attributes that constrain numerical values, physical units and data types. Information about a data set, such as its chronology, purpose and authorship, often classified as ‘metadata’, is also fully defined and may be embedded as an integral part of the CIF. Method attributes are under active development that will permit missing data to be reconstituted on demand where a data block contains other information from which the required data may be deduced (Spadaccini, Hall & Castleden, 2000).

A clear advantage of recording experimental data sets, rich metadata and derived results in the same file, and expressed in a common machine-readable format, is that the results may easily be validated against the source data. Software may check the result set for internal consistency and for consistency with the raw data; and where limitations or potential errors in the results are found, clues may be immediately sought from the supporting metadata (in effect, the accompanying laboratory notebook). IUCr journals take advantage of this ability in the submission and review procedures for crystal structure reports, as explained in the next section.

Similar issues are addressed in principle by much of the current research work in the XML world on schemas, metadata standards and common object standards (World Wide Web Consortium, n.d.). In practice, relatively few production systems have yet evolved to take full advantage of these ideas. In contrast, the IUCr has been using CIF efficiently in journal production for almost a decade. There is no doubt, however, that work will be needed in due course to integrate CIF more completely into standard XML-based publication systems.

Taken together, the components of CIF **-** its file syntax, public dictionaries, extensibility, dynamic typing and relational declarations **-** form a rich domain ontology that is lacking in many other computational environments. Developments in the understanding of document object models will facilitate the interchange between structured document formats when and as practical requirements demand (Murray-Rust, 1998).

5 THE CIF PUBLICATION CYCLE

The IUCr journals that report crystal structure determinations (*Acta Crystallographica Section C* and *Section E*) provide
an excellent example of integration between data and publication, between experiment, analysis and data reuse.

5.1 The authoring phase

Initially CIF was developed to facilitate analysis of data from different diffractometers. In consequence it is possible to generate the file that will eventually be submitted for publication directly from the experimental hardware. The raw data are imported into structure solution and refinement programs, and at the end of many cycles of refinement the programs output a complete description of the data processing, analysis and determined structure. The author then annotates the file with the discursive text of the article and any additional information required by the journal. No hand processing of the data content is required. This fact alone goes a long way towards eliminating transcription errors.

Because a CIF is an ASCII text file, an author may use standard operating-system editors or word processors. However, this reintroduces the possibility of inadvertent modification of content, and various strategies exist to counter this. Some authors edit the text of their article in a separate CIF from that containing the data, and merge the two through software applications or simple file concatenation. A more sophisticated editor with a graphical user interface is currently under development at the Cambridge Crystallographic Data Centre.

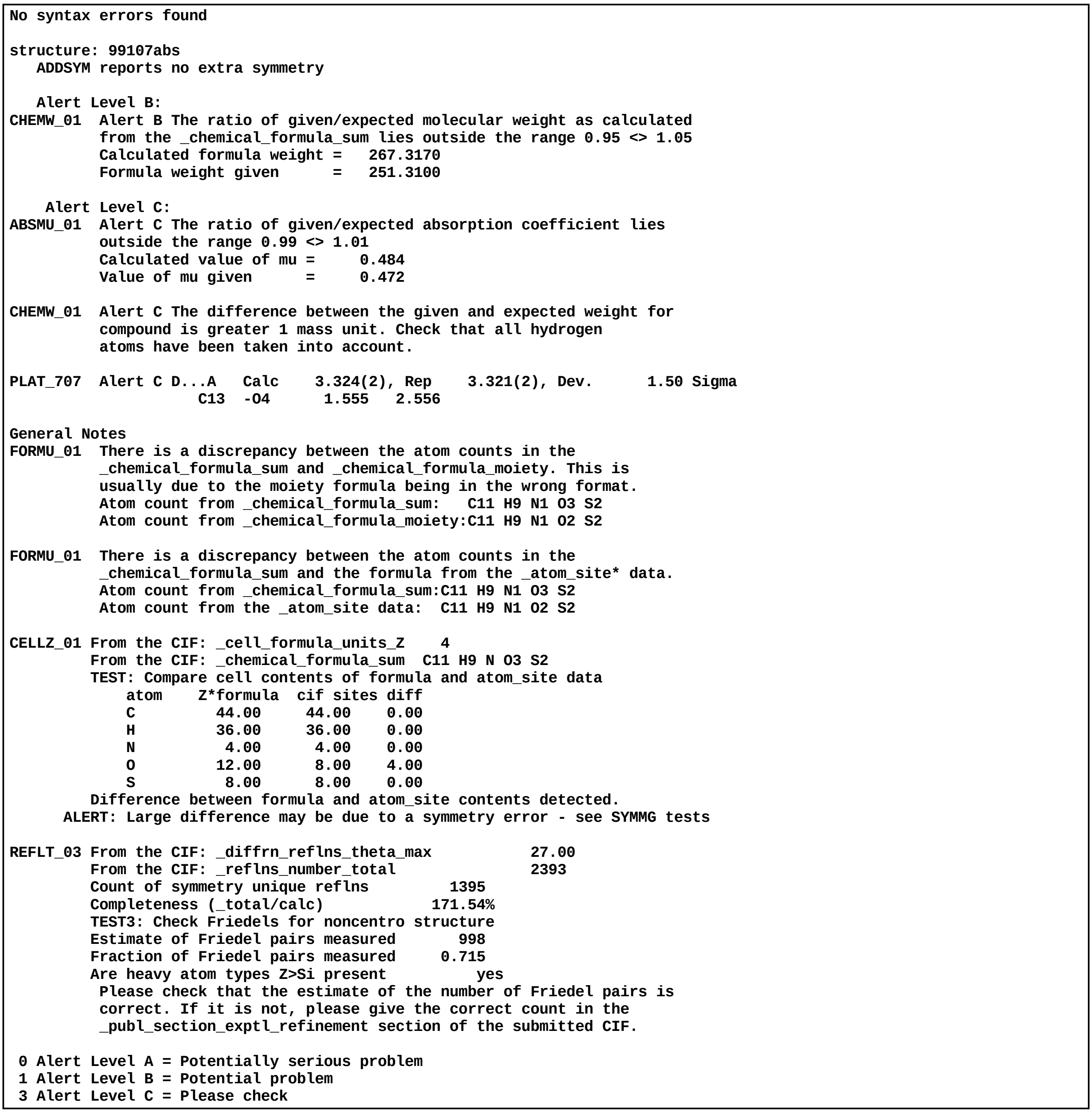
5.2 Submission and validation

Tools are available from the IUCr web site to help a prospective author to check the syntactic integrity and completeness of the submission, and also the consistency of the crystal structure that is being reported.

Templates are available to indicate mandatory data items and those that are optional but desired. Software is available for an author to check the presence or absence of mandatory items, spelling of data names, and the conformance of data values to the constraints specified in the standard data dictionaries.

The author may then submit the CIF, *via* email or a Web interface, to a more rigorous data validation service (*checkCIF*) that runs on a server at the IUCr editorial office (International Union of Crystallography, 2002a). This server first runs file sanity checks, and rejects the CIF outright if it fails any. Where the CIF is intact and complete, the server then extracts the crystallographic and chemical structural data, and validates them against a number of algorithms designed to explore the reasonableness and consistency of the structural model (International Union of Crystallography, 2002b). The data validation criteria form an important part of the journals' acceptance criteria. Many of the checks are performed by a package available separately from the University of Utrecht (Spek, 2000).

```
checkCIF Results for CIF (example_vrf.cif)
        validation tests (Version 0.9) 6th April 2001
```



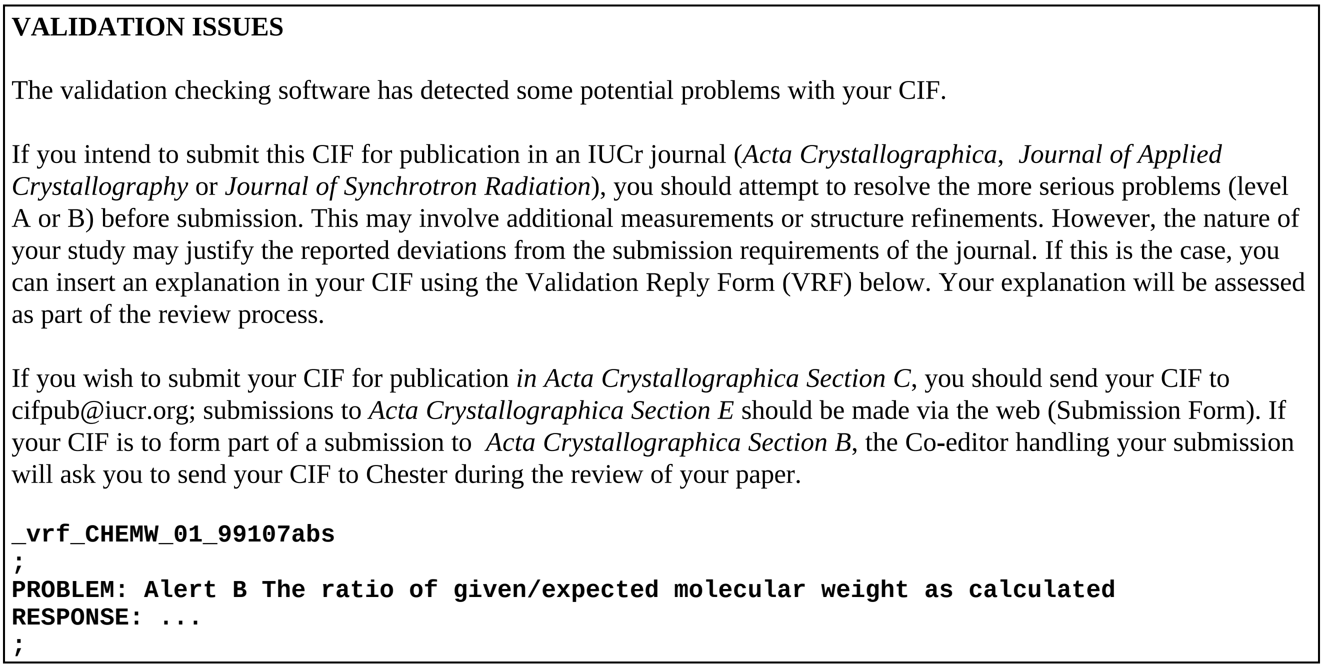
```
No syntax errors found

structure: 99107abs
   ADDSYM reports no extra symmetry

   Alert Level B:
CHEMW_01  Alert B The ratio of given/expected molecular weight as calculated
          from the _chemical_formula_sum lies outside the range 0.95 <> 1.05
          Calculated formula weight =   267.3170
          Formula weight given      =   251.3100

    Alert Level C:
ABSMU_01  Alert C The ratio of given/expected absorption coefficient lies
          outside the range 0.99 <> 1.01
          Calculated value of mu =     0.484
          Value of mu given      =     0.472

CHEMW_01  Alert C The difference between the given and expected weight for
          compound is greater 1 mass unit. Check that all hydrogen
          atoms have been taken into account.

PLAT_707  Alert C D...A   Calc    3.324(2), Rep    3.321(2), Dev.      1.50 Sigma
                     C13  -O4      1.555   2.556

General Notes
FORMU_01  There is a discrepancy between the atom counts in the
          _chemical_formula_sum and _chemical_formula_moiety. This is
          usually due to the moiety formula being in the wrong format.
          Atom count from _chemical_formula_sum:   C11 H9 N1 O3 S2
          Atom count from _chemical_formula_moiety:C11 H9 N1 O2 S2

FORMU_01  There is a discrepancy between the atom counts in the
          _chemical_formula_sum and the formula from the _atom_site* data.
          Atom count from _chemical_formula_sum:C11 H9 N1 O3 S2
          Atom count from the _atom_site data:  C11 H9 N1 O2 S2

CELLZ_01 From the CIF: _cell_formula_units_Z    4
         From the CIF: _chemical_formula_sum  C11 H9 N O3 S2
         TEST: Compare cell contents of formula and atom_site data
             atom    Z*formula  cif sites diff
             C         44.00     44.00    0.00
             H         36.00     36.00    0.00
             N          4.00      4.00    0.00
             O         12.00      8.00    4.00
             S          8.00      8.00    0.00
         Difference between formula and atom_site contents detected.
      ALERT: Large difference may be due to a symmetry error - see SYMMG tests

REFLT_03 From the CIF: _diffrn_reflns_theta_max           27.00
         From the CIF: _reflns_number_total               2393
         Count of symmetry unique reflns         1395
         Completeness (_total/calc)            171.54%
         TEST3: Check Friedels for noncentro structure
         Estimate of Friedel pairs measured       998
         Fraction of Friedel pairs measured     0.715
         Are heavy atom types Z>Si present          yes
          Please check that the estimate of the number of Friedel pairs is
          correct. If it is not, please give the correct count in the
          _publ_section_exptl_refinement section of the submitted CIF.

 0 Alert Level A = Potentially serious problem
 1 Alert Level B = Potential problem
 3 Alert Level C = Please check
```

Figure 4. Example *checkCIF* output indicating a small number of potential problems in a CIF submission.

Typical of the consistency checks that are run are: comparisons of listed positional coordinates of atoms with reported geometric bonds, angles, torsion angles and intra- or intermolecular contacts; chemical formulae derived from the coordinate listings and supplied by the author; agreement between reported and derived molecular weights; agreement between calculated and reported absorption coefficients; compatibility of the reported crystallographic space group with the atomic positions.

Figure 4 is an example of the output generated by a submission with a small number of suspect features.

Where the validation software detects apparent gross errors or inconsistencies, or data values that are outliers to expected distributions, a *pro forma* text form is produced appropriate to the individual article, which the author must complete. This *validation report form* is incorporated into the CIF that is eventually submitted for publication. Figure 5 gives an example of such a form.



VALIDATION ISSUES

The validation checking software has detected some potential problems with your CIF.

If you intend to submit this CIF for publication in an IUCr journal (*Acta Crystallographica*, *Journal of Applied Crystallography* or *Journal of Synchrotron Radiation*), you should attempt to resolve the more serious problems (level A or B) before submission. This may involve additional measurements or structure refinements. However, the nature of your study may justify the reported deviations from the submission requirements of the journal. If this is the case, you can insert an explanation in your CIF using the Validation Reply Form (VRF) below. Your explanation will be assessed as part of the review process.

If you wish to submit your CIF for publication *in Acta Crystallographica Section C*, you should send your CIF to cifpub@iucr.org; submissions to *Acta Crystallographica Section E* should be made via the web (Submission Form). If your CIF is to form part of a submission to *Acta Crystallographica Section B*, the Co-editor handling your submission will ask you to send your CIF to Chester during the review of your paper.

```
_vrf_CHEMW_01_99107abs
;
PROBLEM: Alert B The ratio of given/expected molecular weight as calculated
RESPONSE: ...
;
```

Figure 5. The validation report form (the portion in fixed-width typeface) returned as part of the *checkCIF* output to the prospective author of the example in Figure 4. The author must check and modify any erroneous numerical values in the CIF before resubmission. Alternatively, the data item supplied may be cut and pasted into the CIF, and edited to explain the reason for the persistent error.

The different journals have different acceptance criteria related to this validation report form. For *Acta Crystallographica Section C*, a paper will be rejected if it does not adequately explain the problems identified by the checking software. *Section E* is a little more lenient, in effect permitting the publication of structures of lower quality provided the reviewers are convinced that the article is of real scientific interest. In any case, problems identified by *checkCIF* are reported in a supplementary document accompanying any paper published in *Section E* so that subsequent researchers are fully aware of any limitations of the reported structure.

As part of the *checkCIF* process, the author may also 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.

Authors are *required* to run *checkCIF* prior to the formal submission of their article. However, they are at liberty to use it at an earlier stage when still contemplating submission, and experience suggests that many authors value it as a diagnostic tool, and one that provides early indications of shortcomings that referees might challenge.

5.3 Proofing and reviewing

The CIF submitted for publication contains discrete data fields in a rather arbitrary order, and comprises only ASCII characters. It is transformed into a more conventional article through a number of filters. The first produces a TeX file that generates the version of the article seen by the referees. A companion service to *checkCIF*, known as *printCIF*, is available *via* email and the Web for authors to preview their submissions as formatted by this software (International Union of Crystallography, 2002c).

The formatted document and a set of validation and checking reports (including the *checkCIF* output and some additional diagnostic materials) are sent by the journal editorial staff to referees. The objective is to present the referee with a complete understanding of the technical quality of the structure determination, so that recommendation for publication can be made according to more general criteria.

If revisions are required from the author, a new CIF must be generated and again cycled through the validation process.

5.4 Publication

When the article is accepted for publication, the final version of the CIF is processed by the TeX filter previously mentioned and then by a second filter to generate SGML suitable for typesetting and HTML generation. For *Acta Crystallographica Section E*, which is a purely online journal, articles are published on the Web as soon as all editorial processing has been completed. Typically articles pass through the entire refereeing process and are published within a month of submission. The traditional periodical gathering of articles into monthly issues and annual volumes is retained.

For convenience, the formatted article contains only a selection of the material contained in the submitted CIF: typically the experimental conditions and crystal cell parameters, a subset of interesting geometrical features and the scientific discussion. Readers requiring the complete data set may download a fuller set of data in CIF format by selecting appropriate hyperlinks. These may include listings of diffraction peaks or powder profiles; or the complete set of atomic coordinates, atomic displacement parameters and geometry. In the latter case, separate links are provided that instruct the server to serve the files with different MIME content-type headers so that a reader may direct them to different applications. Typically one link may use MIME type 'text/plain' so that the file may be read in the browser or a text editor, while another link uses 'chemical/x-cif' allowing coupling to a molecular graphics visualiser (Figure 2).

Given the ubiquity of CIF-capable software in crystallography, it is quite feasible for a reader to import the diffraction data into local programs to re-refine a structure or to use it as the seed for solving a related one.

5.5 Graphics

At present the CIF format is not well suited to handling graphics files for publication purposes, and illustrations to a CIF-generated article must still be processed manually. There are a number of developments which will address this problem. An 'image CIF' extension has been devised, initially to store and exchange image plate data sets (International Union of Crystallography, 2000). In principle it could be used to transport arbitrary graphics files. Secondly, proposals are under consideration for encoding standard graphics formats as ASCII text streams and embedding these as suitably tagged data fields within CIFs. It is likely that the embedded fields would contain information about content type similar to the MIME conventions, so that it would be easy to invoke handlers for different types of content. A third possibility, appropriate especially for standard representations of molecular structures and crystal packing diagrams, is to design a set of generic commands expressible as CIF data items and suitable for generating, sizing and orienting desired views. A final, and rather speculative, idea is the creation of a separate graphics language in CIF format. It is difficult to justify this last approach in practical terms, given the existing wide variety of graphics languages; but the capability certainly exists.

In the longer term, an approach that is especially promising is stylesheet-driven transformation of subsets of the data in the file to an appropriate graphical display format. An example of how this might be done for chemical information files where initial XML data is transformed into the scalable vector graphics (SVG) language for visualisation is provided by Gkoutos, Murray-Rust, Rzepa, Viravaidya & Wright (2001). Similar techniques could be applied to CIF data (either through an intermediary XML transformation or direct from the native file format).

Such techniques are especially well suited to graphical representations of the data content of a file (as distinct from illustrative graphics relating to other material). Obvious applications would be to chemical structural diagrams or powder diffraction profiles in two dimensions, or molecular models, atomic displacement plots or crystal packing arrangements in three dimensions. Such visualisations may take place with interactive software that allows a user to rotate, scale, crop or otherwise manipulate the graphic. This can certainly enhance the value of a graphical rendering, allowing the user to gain a deeper understanding than would be possible from a static image. On the other hand, publication illustrations are often sized, cropped or oriented to emphasise a particular feature, and so a full graphics solution must allow for author-created views to coexist alongside the user's freedom to establish and explore other views.

5.6 Validation services for other publishers

The IUCr uses the automated *checkCIF* procedures to assess the technical quality of reported crystal structures as an essential element of the peer review process. Other publishers may exercise less stringent acceptance criteria. The IUCr has therefore developed, in cooperation with a number of publishers of chemical journals that report crystallographic information, variants of *checkCIF* accessible *via* the Web by prospective authors of those journals. In line with the policies of individual publishers, these variants are mostly of an advisory nature. However, in encouraging authors to use such automated validation services, the objective is to improve the quality of crystal structure reports deposited as part of the scientific literature.

Contemporary developments in crystallographic hardware and software greatly facilitate the determination of crystal structures; but they also facilitate error by inexperienced or occasional practitioners. The ability to validate structures during the refinement process, either by use of a public *checkCIF* facility or by implementation of local checking software, then becomes a valuable teaching aid.

6 INFORMAL PUBLICATIONS

The academic journals described at length above represent the formal research publications of the IUCr. In addressing the needs of its community for education, communication and awareness, the IUCr has developed the content of ‘Crystallography Online’ (International Union of Crystallography, 2002d) into a powerful and comprehensive information resource that fully complements the online journals resource.

The site is structured under a number of logical headings, designed to reflect the various topics of interest and relevance to the community of crystallographers. The main topic headings are:

- Crystallographers Online
- Crystallography News Online
- Crystallographic Organizations Online
- Crystallographic Resources and Information Online
- Crystallographic Activities Online
- Crystallographic Education Online
- Crystallography Journals Online

Care has been taken to provide rich links between subtopics, so that users can both locate their logical position within this topic hierarchy and navigate speedily to other topics of interest.

Among the useful news-based resources are meeting calendars and employment listings. These extend and update the meetings and employment notices carried in the print editions of the journals.

There are also: obituaries and notes about personal awards, achievements and appointments; directories of people, databases, laboratories and research facilities; an online version of the IUCr’s print Newsletter; details of the work of the IUCr and its commissions; discussion lists; educational resources.

Online services allow authors of articles in the journal to download proofs and reprints, and to check the status of their submissions in the editorial process.

7 FUTURE DIRECTIONS

A number of developments are in hand or planned to consolidate the philosophy of integration that underpins the IUCr’s publishing operations.

Individuals may make use of the Web resources in many roles: as readers of the journals, authors, subscribers, referees or editors. Editors may wish to locate suitable referees for an article; readers may wish to discuss the content of an article, receive e-mail alerts of new issues, or communicate with editors. To facilitate these multiple interactions, and especially to permit unique identification of an individual contributing to the information pool, the existing contacts database will be extended and will form the basis of a registration and authentication system for participating users.

Further progress will be made in extending the links from journal articles to structural and chemical databases, and in encouraging authors to supplement their articles with suitable multimedia and supporting data files.

Additional structure will be introduced to improve searching of the journals and other information sources.

Work is needed on defining semantic transformations between CIF and XML data representations, both to improve the integration between the two production systems used in-house, and to enable interoperability between crystallographic and other data standards.

Because crystallography is a relatively small and well-defined field, it has been possible to establish an integrated information resource that serves the community very well. However, based on standards and open, widely-available software tools, the systems that have been built are extensible and compatible with other well-structured systems. Integration into a mature knowledge exchange system across the sciences is the long-term goal.

8 REFERENCES

Bernstein + Sons (2000) Homepage of the *OpenRasMol* project. Available from: http://www.OpenRasMol.org

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. & Fitzgerald, P. M. D. (1997) The Macromolecular CIF Dictionary. *Methods Enzymol. 277*, 571-590.

Gkoutos, G. V., Murray-Rust, P., Rzepa, H. S., Viravaidya, C. & Wright, M. (2001) The Application of XML Languages for Integrating Molecular Resources. *Internet J. Chem*., *4*, Article 12.

Goldfarb, C. F. (1990) *The SGML Handbook*. Oxford: Oxford University Press.

Hall, S. R. (1991) The STAR File: A New Format for Electronic Data Transfer and Archiving. *J. Chem. Inf. Comput. Sci. 31*, 326-333.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991) The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography, *Acta Crystallogr., Sect. A*, *A47*, 655-685.

Hall, S. R. & Spadaccini, N. (1994) The STAR File: Detailed Specifications. *J. Chem. Inf. Comput. Sci. 34*, 505-508.

Higher Education Digitisation Service (n.d.) Homepage of the Higher Education Digitisation Service. Available from: http://heds.herts.ac.uk.

International Union of Crystallography (2000) imgCIF dictionary. Retrieved April 22, 2002 from the IUCr Web site: http://www.iucr.org/iucr-top/cif/imgcif/index.html

International Union of Crystallography (2002a) *checkCIF*. Retrieved April 22, 2002 from the IUCr Online Journals Web site: http://journals.iucr.org/services/cif/checking/checkform.html

International Union of Crystallography (2002b). IUCr data-validation tests. Retrieved April 22, 2002 from the IUCr Online Journals Web site: http://journals.iucr.org/services/cif /datavalidation.html

International Union of Crystallography (2002c) *printCIF*. Retrieved April 22, 2002 from the IUCr Online Journals Web site: http://journals.iucr.org/services/cif/checking/printform.html

International Union of Crystallography (2002d) Crystallography Online. Retrieved April 22, 2002 from the IUCr Web site: http://www.iucr.org/cww-top/crystal.index.html

ISO (1986) *Information Processing - Text and Office Systems - Standard Generalized Markup Language.* ISO Standard ISO 8879.

Knuth, D. E. (1984) *The TeXbook.* Reading, MA: Addison-Wesley.

Murray-Rust, P. (1998) The Globalization of Crystallographic Knowledge. *Acta Crystallogr., Sect. D. D54*, 1065-1070.

Murray-Rust, P. & Rzepa, H. S. (1999) Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci. 39*, 928-942.

Research Collaboratory for Structural Biology (n.d.) Homepage of the Protein Data Bank. Available from: http://www.pdb.org.

Rzepa, H., Murray-Rust, P. & Whitaker, B. (1998) The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World-Wide Web Information Exchange. *J. Chem. Inf. Comput. Sci. 38,* 976-982.

Spadaccini, N., Hall, S. R. & Castleden, I. R. (2000) Relational Expressions in STAR File Dictionaries. *J. Chem. Inf. Comput. Sci. 40,* 1289-1301.

Spek, A. L. (2000) *PLATON. A Multipurpose Crystallographic Tool.* Utrecht University, Utrecht, The Netherlands.

World Wide Web Consortium (n.d.) Homepage of the World Wide Web Consortium. Available from: http://www.w3c.org