Computerization of the IUCr Editorial Office, Chester: a Review of 1992

The Technical Editor's Office, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England

December 8, 1992

Abstract

The major change felt in the Editorial Office in the course of 1992 has been the transition to full production mode of CIF-based structural papers in Section C of *Acta Crystallographica*. Networking connections have been extended to sites supporting the IP protocol. Some additional hardware has been acquired. Early development work has been initiated with the General Editor of the Ninth Edition of the *World Directory for Crystallographers*.

I. Introduction

This is the third in a series of informal reports on the innovations within the IUCr Editorial Office which are designed to enhance the publishing activities of the Union through modern computer-based practices. It surveys the developments that have occurred since the previous report. The most important of these has been the introduction of 'New Format' papers to *Acta Cryst.* C. These will rapidly replace the old categories of 'Full Article' and 'Short Format', and will become the standard method of reporting crystal structure determinations within that journal. The numeric data in these papers are stored in CIF form, either in files created manually in Chester, or in full electronic transmissions from the authors. The previous report in this series anticipated this transition with cautious optimism; it was impossible to foretell how many submissions would be received in this form, and how robust the typesetting software would prove to be in the face of submissions from untutored authors. In fact, the flow of author-supplied CIFs has risen rapidly (such submissions accounting for about a quarter of all new papers submitted through 1992), and the translation to print of these papers has progressed remarkably smoothly.

Use of CIFs as the storage mechanism for structural papers has as its corollary the building up of a data bank of crystallographic information in a much more complete form than is usual in crystallographic databases. This will have long-term implications for the storage and dissemination of this information. Some pilot experiments have been conducted on a small scale to investigate possible retrieval mechanisms and user interfaces to the data bank. Several of the tools available on the global Internet network may be relevant in this context, and the connection of the Editorial Office to the Internet in summer of this year has greatly broadened the scope of our investigations into this subject.

The storage and distribution of data are problems that also arise in the new *World Directory of Crystallographers* project, and this may well provide a useful test bed for the procedures that will eventually be extended to the structural data currently being archived. Work in this area is being undertaken in collaboration with the General Editor of the *World Directory* project (Professor Yves Epelboin) and with the author of the STAR concept underlying CIF (Professor Syd Hall). In addition to the database aspect of the *World Directory*, the publication itself will be formatted using the *ciftex* tools developed for structural papers, a satisfying convergence of superficially distinct aspects of the Union's publishing activities.

These areas will be discussed in greater detail below, and, as usual, some pointers will be given to the directions of future developments.

II. Structural papers in *Acta*

Two major changes were introduced to the handling of papers submitted to *Acta* C from the beginning of 1992. One change, administrative in nature, required authors to send their manuscripts direct to the Editorial Office, where they would be checked for numeric accuracy and self-consistency before being sent to a Co-editor for review. The other change was to the format of presentation of the paper. Authors were required to supply the basic experimental details on a standard form and to confine their discussion to an Abstract and a Comment section, in order to enable easier keyboarding of the paper as a CIF; or they were invited to send a complete CIF, generated manually or by a computer program, which might also contain the material required for deposit at the British Library.

Such major changes must have come as a shock to many authors, who had over time become used to the standard format required by *Acta*, with its strict adherence to rules of presentation laid down by the Commission on Journals. The announcement of the new procedures appeared in the November 1991 issue of the journal, giving a rather short lead time for authors to adapt to the changes! It was, of course, inevitable that there would be a lengthy overlap period while papers in the old formats that were already in production, or already written by authors, would travel through the system. Nevertheless, in 1992 one in every eleven of the papers published in *Acta* C were in the new format (a total of 83 out of 906). Of these 83, 47 were complete CIFs submitted by authors *via* e-mail or on diskette. More strikingly, of over 700 papers submitted to *Acta* C during the year to date, over 190 (or about 26%) are full author-supplied CIFs. Many of these were created manually or by authors' locally written software, but the proportion generated by standard structure packages is steadily rising. It is our practice to advise authors of structural and syntactic errors in their submissions at an early stage (usually when the submission is acknowledged) in an effort to increase their understanding of the file structure. This approach appears to have paid dividends; many authors have supplied much improved later submissions. Although we do not monitor the author profile closely,

it appears that authors who submit a trial CIF are, in general, likely to submit further papers as CIFs.

The conclusion then must be that the CIF route has proved fairly popular for authors in its first year. Such submissions were accorded priority of editorial attention (necessary for the development of our handling procedures), and we expect that authors who have assisted with these early trials will have been rewarded by fast publication times. There is no doubt that a well constructed CIF can be handled relatively efficiently, and it will be beneficial to the Editorial Office if the proportion of author-submitted CIFs increases. This is likely to be the case as Professor Sheldrick's program *SHELXL*-92 is distributed.

The magnitude of the changes to the editorial process that have occurred with CIF-based publication is so great that it is difficult to assess the true impact on the efficiency of the publishing operation. A complete cost-benefits analysis would need to consider keyboarding costs in Chester, the cost (and benefit!) of complete data checking, demands on additional staff time, as well as the clear financial savings in typesetting costs. Nevertheless, the general picture seems to be that CIF-based papers are, on the whole, no more difficult to process than conventional papers, while author-supplied full CIFs are easier to handle.

The program *ciftex*, described in an earlier report, has matured and is capable of greater functionality than was originally envisaged. Additional pre- and post-processing software combine with *ciftex* to make an impressive tool for translating the textual matter in a CIF to type. Little progress has been made this year on the electronic incorporation of graphical material in the paper (artwork is still photographed and stripped into place during manual page makeup in the traditional printer's workshop), but it hoped that trials can begin in earnest next year to assess the most efficient way of handling artwork.

Data checking

The checking of structural data was described in previous reports, and is now applied to all reported structures in *Acta*. No new tools have been acquired recently (although the new release of *Xtal*, version 3.2, is installed). However, there have been instances when discrepancies between programs have improved our understanding of the assumptions and methods of the different packages, and the use of so many packages in parallel allows for a fruitful interchange of such information. A forthcoming letter in the February 1993 issue of *J. Appl. Cryst.* (Muir & Mallinson, 1993) arose from scrutiny of the ways in which geometric e.s.d.s are calculated in non-orthogonal space groups. The power of the program *MISSYM* to detect space-group errors has been demonstrated in correspondence with its author, Professor Le Page: in all recent cases of space-group errors in *Acta*, *MISSYM* would have indicated the error; the papers concerned (save one) had not been checked in Chester.

Other sections of the journal

There is still no major progress on automating publication of papers in other sections of *Acta* and *Journal of Applied Crystallography*. Several papers have now been published in the journals from TEX source, but so far each has had to be treated as an individual case. Work will need to be done to write a series of macros for TEX which are tailored to our house style; these may be distributed to authors to ensure that they submit their papers in the correct format. Papers produced by various word processors have been imported to *The Publisher* (a powerful TEX-based typesetting package). In both cases, papers can be produced without excessive difficulty; however, the procedures are relatively time-consuming. It will be important to seek ways of maximising the efficiency of handling other types of electronic submission.

In 1992, 57 pages of *Acta* A and 20 pages of *JAC* were typeset in-house from TEX or other electronic input; this represents 6% and 5% respectively of the pages in those journals (scientific papers only; indexes, advertisements and some other announcements are also produced in-house). The proportion is thus small, but not infinitesimal. The trend is towards a gradual increase in numbers, and a large proportion of the papers in the inaugural issue of Section D were produced in this way.

III. Networking

At the end of 1991, the JANET X.25 network was modified to allow transmission of datagram packets conforming to the IP protocol, and thus allowing direct communication with sites on the global IP-based Internet. We modified our communications configuration in late August and have enjoyed full Internet connectivity since that date, although we have not yet applied for delegation of zone authority in the Domain Name Service (what this means is that other sites are not yet able to access us by name, although connection can be established if the other site knows our numeric address).

An immediate benefit of Internet connectivity is the availability of 'anonymous ftp', a file transfer protocol permitting public access to a data collection to browse the directory tree and download (or upload) files of interest. This has allowed us to retrieve many useful tools from public archives (a HPGL-to-PostScript graphics format translator, for example); and it also permits retrieval of deposited material from the Protein Data Bank at Brookhaven. It is likely that this latter procedure will replace the microfiche provision of PDB data that we currently offer.

The anonymous-ftp protocol is moderately secure (if properly implemented), and would provide one way to retrieve CIF data sets (either the archived master files or subsets of data from these files, as policy dictates). The current implementation requests users who log in to supply their e-mail addresses as the login 'password', but validation is limited to the address of the machine that establishes the connection, and not the users. File transfer can be audited by calling machine address, filenames and file sizes. There is no way to permit or bar various subsets of users. (However, the overlying ftp protocol can support individual accounts. The avowed purpose of *anonymous* ftp is to provide universal access.)

Further mechanisms exist in the public domain for the free transfer of information between Internet sites. Some of these may be modified to permit access validation and cost charging, but it should be emphasised that the spirit underlying most of these processes is one of *free* availability of information. The IUCr has, of

course, the conflicting requirements of making available the scientific data it publishes to the full scientific community, yet requiring to raise income to support its publishing operation. At present, a large proportion of this income is derived from journal subscriptions. It is therefore realised that the procedures discussed here may not therefore be ideal (or even suitable) for providing online access.

One such initiative is *WAIS* (Wide-Area Information Server). In its simplest form, a host computer constructs an indexed file of all words in all documents offered for distribution. A search for a word provides the requestor with a list of documents containing that word, ranked as to supposed suitability according to some heuristic figure of merit. The requestor can select one or more of the documents offered, which are then copied to his computer by file transfer. In our context, the 'documents' on offer could be complete CIFs or the textual portions only of structural papers (perhaps more useful for text-based searching). In a more sophisticated implementation, Boolean searches can be made on a set of strings, and context scoring be assigned, so that the information server is asked to provide documents where a given word or phrase appears 'in a similar way' to its use in a previously retrieved reference document.

A related server is *gopher*, which provides a set of menus allowing a user to navigate a set of 'logical directories' in search of information that he requires. The 'directories' may be directories of files on any machine on the network (the precise location is not indicated to the user); or they may be lists of documents returned from a *WAIS* search, or they may indeed be telnet connections to other information systems. The intention is that a user may obtain the information he or she requires by selecting a logical chain of pointers in the available menus, without needing to know anything about the location or type of data stored.

Both these services (*WAIS* and *gopher*) use a client–server model, where responses from (sometimes a large number of) servers are presented to the user as though by a local application.

A third such scheme, also known to us but not yet investigated, is *WWW* ('World-Wide Web'), a protocol established initially within the high-energy physics community that is well suited for transmitting formatted documents (using a variant of SGML, the Structured Generalized Markup Language standard).

Again we emphasise that none of these may be suitable for supplying crystallographic data to the world at large; on the other hand, they are systems that are already implemented, are growing in robustness and functionality, and are familiar to a large community. They may also (as in the case of *WAIS*) be adaptable to transactional logging and costing.

Another point that we must also bear in mind is that not every site will have Internet access (though it is the largest academic network, and continues to grow rapidly), so that there will need to remain methods for supplying data in electronic form *via* e-mail (and probably also on tape/diskette).

IV. Office hardware

The period under review saw the acquisition of two further Sun workstations and an extra 1.3 Gbyte of disk storage. The workstations (SPARC IPCs) are machines of the same power as the three SPARC 1+ boxes originally purchased, but cost less, in following the general trend of continually falling prices for hardware of a fixed capability. They were needed because the increasing proportion of electronic work requires more of the staff to have access to a graphics workstation at any one time. The extra disk space was needed to provide more working space. Checking and typesetting of a paper can each generate files amounting to a Megabyte in total, and it is convenient to be able to keep such files online for the duration of the publishing process. Appropriate partitioning of the new disk also permits the most important files to be backed up to disk automatically every day (this is a further level of security, as daily tape backups are also made of the important filesystems).

The hardware continues to operate reliably, with no major outages during the year. The 3B2/500 continues to function with the greatest reliability, but this machine is now four years old (!) and some thought needs to be given to ways in which its functioning can be taken over by the Sun-based network in the future. The problem is not pressing, but the journals database software on the 3B2 is not portable to the Suns, and so any database tools chosen for the Sun network should ideally have the capability of also handling the type of data currently managed by the 3B2.

The acquisition of high-performance Apple machines (a Quadra 700 'workstation' and PowerBook 170 portable) by the Executive Secretariat has made available to the Technical Editor's office a Macintosh LC, which is attached to the network for data transfer.

V. *World Directory*

As mentioned in the *Introduction*, work is being conducted with Professor Epelboin and Professor Hall to establish a database for the Ninth Edition of the *World Directory of Crystallographers*. The data for this edition will be collected in STAR file format. It is currently an open question whether the database will be accessible via the *Star_Base* software under development at University of Western Australia as a STAR file, or whether more conventional relational database techniques will be used, based perhaps on a commercial product. In either case, this development work is seen as providing essential experience for the larger question of maximising the usefulness to the crystallographic community of the growing archive of CIFs.

The *Directory* itself will be published from the STAR files used to collect and transmit the data. The program *ciftex* will handle the typesetting. Some new TEX macros will be written to format the entries in a typographically pleasing way. It is not expected that the *ciftex* program itself will need any large modifications to handle this particular application.

References

MUIR, K. W. & MALLINSON, P. R. (1993) *J. Appl. Cryst.* In the press.