Computerization of the IUCr Editorial Office, Chester: IV. Network Publishing

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

15 July 1993

Abstract

An account is given of recent additions to the hardware and software inventory in Chester, and of developments in the CIF sphere. A scheme is described for supplying CIFs across the academic networks, and a proposal advanced for publishing the crystallographic data in the growing CIF archive that may raise revenue..

Introduction

This report is the fourth in an occasional series of papers prepared by the Technical Editor's office in Chester to describe the process of automating journal production. Previous reports in the series are referred to within this document as: **I** (*Computerization of the IUCr Editorial Office, Chester: a Progress Report,* dated 19 October 1990), **II** (*Computerization of the IUCr Editorial Office, Chester: a Review of 1991*, dated March 4 1992), and **III** (*Computerization of the IUCr Editorial Office, Chester: a Review of 1992*, dated December 8 1992). The current paper reports some recent developments and considers some technical and policy matters relating to the provision of journal papers across the research networks.

Developments since the previous report

Paper **III** in this series described the office equipment and software inventory at the end of 1992. In spring 1993 a portable SPARCbook computer was acquired. This has the processing power of a SPARCstation II desktop workstation, but is fully portable, and compatible with all Unix and Sun applications that do not involve the old Sunview windowing system. It has 32 MB DRAM memory, and two 125 MB internal hard disks: these will be replaced by 250 MB disks when these become available from the supplier. The system has an Ethernet connection, serial port, SCSI connection, internal fax/modern (though this is not licensed for use on the UK public telephone system), audio I/O port, keyboard/mouse port allowing connection of an external Sun keyboard and/or mouse, and a video port allowing the use of an external VGA or Sun monitor. The system has already been used to demonstrate site software at conferences and other meetings. With the addition of an external Sun monitor and keyboard, it will also become an integral part of the in-house network. We did experience some hardware faults soon after the system was delivered, but it is hoped that these have now been repaired.

A serial connection between this system and the Executive Secretariat's Apple Powerbook permits maximum portability of the office computing environment.

New versions of some checking programs have been installed, including versions of Spek's programs *PLATON* and *PLUTON* for analysing and displaying crystal structures (these programs can input CIFs direct); *SHELXL*-93 and associated utilities (this is the new version of George Sheldrick's refinement program); and *BUILDER*, a new version of the Cambridge checking programs that has a graphical interface. We have also obtained a copy of the *PROCHECK* suite of programs designed to check protein structures [Laskowski, MacArthur, Moss & Thornton (1993). *J. Appl. Cryst.* **26**, 283-291]. The checking routines have been enhanced to allow the generation of *ORTEP*-style thermal ellipsoid plots, which are now supplied to Co-editors to assist in their evaluation of crystal structures.

Table 1 Summary of CIFs in hand at July 13 1993

In total 2417 CIFs are in house, of which 1563 have been archived

Number of centralised submissions:	1269
Number generated in-house:	2057
Number generated by *PARSTCIF*:	31
Number generated by *NRCVAX*:	26
Number generated by *SHELX*:	26
Number generated by *Xtal*:	20
Number generated by *TEXSAN*:	11
Number generated by *PLATON*:	11
Number generated by other:	36
Number with no source given:	199

Of the centralised submissions 360 (28%) are external sources

CIF

The Crystallographic Information File (CIF) appears now to be well established as a means of information interchange in the small-molecule crystallographic community. Table 1 illustrates the number and variety of CIFs on the Chester computer system at the time of writing.

The Technical Editor's office has over the last few years sent representatives to several national and international conferences to explain the use of CIF, and all current *Acta* C papers are typeset, at least in part, from CIFs. The proportion of papers from complete CIFs supplied by authors has for some time hovered in the region of a quarter to a third; it is hoped that this proportion will increase as Sheldrick's program *SHELXL*-93 spreads through the community.

After a period of tranquillity in the area of CIF development, 1993 saw a sudden increase in activity. The definitions for the macromolecular and powder diffraction extensions to CIF were fleshed out, but the requirements of the macromolecular community for explicit data interrelationships within CIF files has led to a revision of the DDL, the dictionary definition language used to present CIF dataname definitions in electronic form. These revisions will have some effect on the CIF Core Dictionary, but the changes that will occur to the Core Dictionary should be transparent to the general user, as they relate only to the presentation of information in the Dictionary, and not to the existing definitions themselves. A small number of revisions to

the Core Dictionary will be considered by the Committee for the maintenance of the CIF Standard (COMCIFS), but the Dictionary as published in *Acta Cryst.* (1991), A**47**, 655–685, has proven very robust to date.

Members of the Technical Editor's staff have been involved with these developments, and have also played a small part in the development of the proposed Molecular Information File (MIF) standard for chemical information. This is a reflection of the fact that the archive of structural data at Chester remains the largest current 'application' of CIF, but it also illustrates the extent to which the publishing activities of the Union are increasingly becoming entwined with more general data management activities. It is this divergence from the traditional skills of typographic representation that will characterise the Union's publishing activities in the future.

Document markup

We have explained in earlier reports (see especially **I**) how the program TEX is used to format the structural CIFs in *Acta* C; it is also used for some papers in other sections of the journals. The TEX source files may be regarded as *marked-up* versions of the original document, where the *markup* is a method of tagging the file to indicate to subsequent applications programs (such as the PostScript output converter or a screen display program) how the document is to be represented in its final form. Many of the TEX instructions included in the files we produce refer to small points of typographic detail (say, a decrease in spacing between the columns of a table to improve its aesthetic appearance or to fit it within a column on the page); as such, these tags are regarded as of poorer style than *structural* tags, which indicate the intention underlying the representation of a portion of the text. Such structural tags may be interpreted differently by different document processing programs, so that a single set of source tags may have use in different classes of application. As a simple example, a phrase in a TEX source file may be rendered in italic type by use of a code such as `\it` which invokes a specific typeface. However, it may be considered better style to use a different code, such as `\em` (for *emphasis*) which renders the phrase in italic type on the printed page, but in inverse video on a computer screen.

Descriptive codes such as these may indeed be defined in the plain TEX environment; in practice, they arise naturally within the different dialect of TEX called LATEX, and so it may be considered that LATEX offers a better environment for tagging documents that may be handled electronically in several different ways. In this respect, a LATEX file may more easily be translated into the international standard SGML format. In consequence, we shall consider carefully the development of a LATEX 'style' (or set of tags constructed by ourselves) which will allow papers to be rendered typographically or in some other style of output format without change to the source file.

There are two aspects to this. It will generally be preferable for prospective authors to compose an electronic submission in LATEX, because the descriptive codes should be more meaningful to the author who is not an expert in typography. So it makes good sense to provide a LATEX style appropriate to the papers in *Acta* sections A, B and D, and *J. Appl. Cryst.* which can be used as a template by intending authors. Secondly, the existing plain TEX macros in structural papers as output by *ciftex* will need to be recast in a LATEX compatible form. In most instances this will not pose any great problem, though there will undoubtedly be significant challenges in devising a robust set of instructions for page makeup.

The aim behind this is to produce SGML-tagged documents which can be typeset or handled in other document management schemes (one such is described in the section on *Internet services*). Why not work with SGML files from the start? That is, why not write a *cifsgml* filter to generate SGML tags from an input CIF? There is no great conceptual barrier to this approach, but we do not know of any public-domain SGML tools which have the same power and widespread utility as TEX (though for a description of an SGML-compatible formatter based on TEX, see the description of *The Publisher* in **I**).

Document delivery

This phrase is increasingly used to refer to the supply of documents by electronic means, ranging from fax transmission of typeset pages through network transmission of ASCII and PostScript files. It is a wide and complex topic, and is relevant to the work of publishers, distributors and libraries. Undoubtedly we must consider the possibility of supplying versions of the printed page by fax or by transmission of scanned page images. It is hoped to produce in due course a study report on the issues involved, and their implications for the IUCr's publishing activities.

It is probable that moving into this area of direct document delivery will involve further significant hardware investment, in scanning equipment, mass storage devices, data format interconversion software and telecommunications facilities. Further, mechanisms to recover the costs of supplying individual documents need to be developed, and will be complex and costly to maintain. This aspect of subscriptions or direct charging should probably be investigated closely in collaboration with Munksgaard.

However, although this is a large area for future development that must be taken seriously, there is already to hand for the case of structural papers an easier route to 'document delivery' — direct provision of CIFs. Already over 1500 CIFs are archived in Chester, with nearly a thousand more in hand. Most of these represent numeric summaries of the structure solution, but more than 350 already contain the full text of papers in *Acta* C, and it is intended that this proportion should rise. Technically, it is very straightforward to make these available across the network by e-mail, anonymous file transfer protocol ('anon-ftp') or other means as described below. The recipient will have the benefit of obtaining usable numeric data in machine-readable form, while the printed page can be regenerated, if necessary. However, attractive though it is to have the proposition of recovering the typographic form of the journal, it must be remembered that the majority of readers are interested, at least for *Acta* C type papers, in having the data in a manipulable form.

What of cost retrieval? We make the suggestion here that CIFs should be distributed across the network *free*

of charge, but subject to a limit on the number that may be obtained in this way. There may, however, be other ways of generating income from this approach, one of which we shall discuss later.

Network connections

The IUCr office is directly linked to the JANET network; which is the UK academic X.25-protocol-based academic net. It is also possible to 'wrap' IP protocol datagrams in X.25 packets, a mechanism which allows direct data transfer across the Internet. There is also essentially direct connection to European X.25 networks. The connection is currently at 9600 bits per second (bps), an adequate speed for supporting a small number of terminal-based sessions or for downloading small to medium-sized files by anon-ftp. The charge for this service is about £ 8000 per year. It would be possible to upgrade to a 64 kbps line for about £ 20 000 per year. JANET is able to support line speeds up to 2 Mbps and will soon be operational as a 34 Mbps network, but the technical and cost implications of acquiring such a high-capacity line must make it very unlikely that we would seek such a service for a considerable time to come. (However, it is likely that the increase in capacity of the network as a whole will reduce traffic congestion even at low speeds.)

It is now possible to obtain direct connections to the Internet through some commercial companies operating in the UK, but the costs of doing this are not much less than the cost of the JANET connection, and the advantages of direct X.25 connectivity and the use of other gateways are lost. It is also possible, in the future, to consider ISDN connections (these are data-carrying connections across standard telephone lines); but while ISDN is a low-cost option for obtaining network services for occasional users, the cost to a service provider such as ourselves of keeping a standard telephone line open to network traffic 24 hours a day again becomes comparable to the standard network charges. Further, ISDN will become a viable option only when a high proportion of sites wishing to connect have the facility to place an ISDN call — something which is not likely to happen for many years yet.

The current situation is therefore satisfactory, in that we can reach and be reached by sites on JANET, European (and Australasian) X.25 networks, and the Intemet. We also enjoy the benefit of mail gateways to BITNET, uucp and other smaller networks, and are therefore in a position to provide services to most of the world.

However, we must remain aware that other parts of the world are not always so favourably endowed. Much of Africa still has no standard network connections; China and other Asian countries have very limited *and very expensive* network access; and India and the former Soviet bloc countries can find it difficult to reach the international networks for reasons of national financial stringency and operational deficiencies. For this reason, any network services we provide should be in parallel with services to provide data by e-mail alone, or even on diskette.

This is one reason why we believe the provision of CIFs over the network should be made available at no cost. So long as we are able to impose a limit on the *number* of requests serviced, a requestor from China or India should be at no financial disadvantage compared with a requestor from the USA, beyond the higher cost in real terms of using the network within the requestor's own country.

It is also important that the use of academic networks not be abused. Many networks will not permit the transfer of data on a saleable basis: their charters declare that they exist for the *free* interchange of academic information. The problem is compounded by the fact that international data transfer can extend over several networks. It is possible (though probably unlikely) that we might be able to convince JANET that we should be able to charge for data sent over JANET; but if the file is sent to France, that would absolutely contravene the conditions of use of the French network. And even if the file is destined for another country, if the data packets are routed through France, one might be in breach of the conditions of use of the bridging network.

Probably none of these objections are insuperable for a purely commercial concern. However, it is our belief that the Union will best serve the interests of its members by supplying CIFs free of charge in network transmission.

We now consider ways in which the CIFs may be made available. Most simply, they may be requested by e-mail, giving the reference number cited in each paper printed in *Acta*. A service permitting this has already been established on a trial basis, though so far there has been minimal interest shown in it. There are probably several reasons for this. First, it is a new service, not yet well publicised, and known to only a few crystallographers. Second, the community is not yet aware of the usefulness of having files in CIF format. Still relatively few applications exist that can handle CIFs directly, while the average CIF is less comfortable to read than the parent journal paper (though one of the design aims of CIF was that it should indeed be readable by humans in a way that older file formats were not). Third, the mechanics of requesting files by e-mail, while open to all users with mail access, are complex compared with the ease of access of interactive services such as those listed below.

We believe that the community will be more interested in downloading CIFs when more tools are available for processing the CIFs, and when it is easier to log in or access the IUCr data archive interactively.

ciftex

One application that we think should be made available to the community at large is the program *ciftex* that we use to produce a TEX file from a CIF. A crude version of this software, called *cifms*, has already been released, and has proved quite popular (about 30 copies have been distributed to date, and many other requests have come in for a DOS or VAX-compatible version). However, while *cifms* does provide an idea of the way in which the journal paper will be constructed, it does not have the ability to display the output in different typefaces and with special typographic symbols. Use of *ciftex* requires that the user have a copy of TEX, but this is not a grave problem since TEX is in the public domain and is already very. widespread. The user (with TEX) will then be able

to obtain an exact reproduction of the typeset paper from the CIF.

It is likely that the majority of potential users would use *ciftex* purely as a visualization tool. It is possible, however, that others would wish to modify it for their own purposes. Where this is for purely private or local use, we can see no problems with this — while *ciftex* is not a very complex program, it could serve as a useful template for developing other CIF software. However, there is of course some concern that other bodies might seek to use the program for commercial gain. We suggest that this should be permitted so long as the copyright in the program remains with the IUCr and acknowledgement is made of the use and provenance of the program. In practice, any commercial use would require the recasting of the TEX macros used to a form consistent with the style of the journal concerned — a non-trivial exercise in itself!

Internet services

Although the Chester office has been connected to the UK academic network for some time (see **II**), our perspective on the use of wide-area academic networks for electronic publishing changed radically when the JANET network was modified to permit direct access to the Internet (the phrase applied to the global collection of networks sharing the (TCP/)IP protocol, and thus able to freely intercommunicate). From its inception, the Internet has grown around the idea of *distributed* resources. For instance, to contact a computer in Australia, my local computer needs to resolve the numeric address of the target computer's name. But my computer does not contain a table of all possible name→address mappings (which would need to be a very large table to accommodate the 1.4 million or more systems currently connected). Instead, it contains a small table of commonly accessed mappings, and a set of pointers to other systems which can resolve some portion of the target address. A standard protocol allows my computer to interrogate these 'name resolvers'; if they are unable to provide the required address, they will in turn interrogate other systems with more specific information on a subset of the possible addresses.

In the same spirit, Internet protocols have been developed to permit the sharing and interchange of other types of information, ranging from the time of day through the identity of users logged on to the system, to more organised data made publicly available by service providers. In our report **III**, we have already referred to a number of these services. We mention again here *gopher*, which is a program that makes available for public retrieval directories and files on a computer, together with some other services (interactive *telnet* sessions, simple database ('phone book') searches, and indexed retrieval of documents containing specified whole words). Computers which are set up to supply these services ('*gopher* servers') can contain pointers to other such servers; these pointers appear to the inquirer simply as additional items in the menus he can access. We have implemented a demonstration *gopher* server on the main IUCr computer (see Figs. 1–3). This allows the retrieval of files of CIF software normally supplied by the *sendcif* information server (or accessible by anonymous ftp), and also retrieval of archive CIFs,

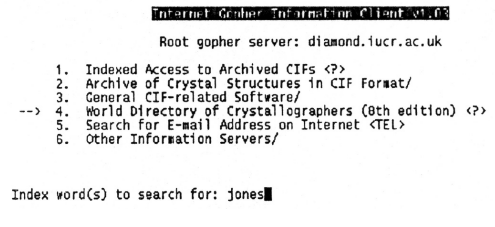

Fig. 1 Typical screen display during an interactive IUCr *gopher* session. Directories of files available for transfer to the user are indicated by a trailing slash after the menu item. *WAIS* databases are indicated by a trailing query mark, and gateways to *telnet* services are labelled by TEL. The user has selected option 4 on the main menu and is prompted to enter a search string. The progress of the session is continued in Figs. 2 and 3.

stored in a separate directory for each issue. It also provides a mechanism for retrieving CIFs from a single-word search (see the discussion of *WAIS* below), and a similar interface to the previous (eighth) edition of the *World Directory of Crystallographers*. There is also a *telnet* gateway to a service which endeavours to find the e-mail address of a user on the Internet (given some seed information), and a set of pointers to *gopher* servers at the Protein Data Bank, Nucleic Acids Data Bank, UK TEX Archive, and some other sites.

The *gopher* protocol assumes a client–server relationship between the sites involved in the transaction. That is, the site supplying the information runs a daemon, or server, program (usually *gopherd* or *gn*), while the inquiring site uses a separate client program (*gopher*, *gopher+*, *TurboGopher* . . .) to handle the information supplied by the information provider. A consequence of this is that the client program can interface with local software. As a specific example, we can consider an example where the IUCr supplies CIFs from its *gopherd* server. A crystallographer with an Internet connection can retrieve CIFs from the IUCr, and read them as plain-text files on his screen. Alternatively, he can directly pipe the received file to a formatting program — say, *ciftex* — and view a complete representation of the *Acta* paper on his screen or laser printer. Or he may pipe the received file direct to *PLUTON* or some similar interactive graphics program, and visualise the molecule directly. Note that neither the server nor the client *gopher* programs need be modified in any way to permit this — the user may configure his *gopher* client to pass the files it receives by the standard *gopher* protocol to the *ciftex* or other local program.

We have already mentioned the use of *WAIS* to provide a mechanism for retrieving documents containing a given word. The *WAIS* (the acronym stands for Wide-Area Information Server) protocol permits interrogation

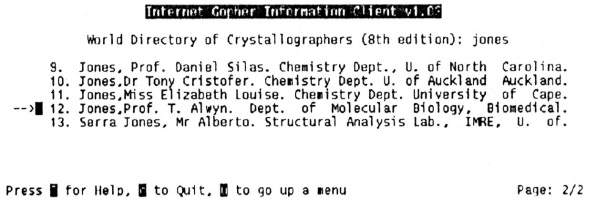

Fig. 2 The WAIS search on 'Jones' returns a list of 13 documents containing the search string. The user advances the screen cursor to item number 12.

```
This section is from the document '//.wdc8/wdc8'.

Jones,Prof. T. Alwyn.  Dept.  of  Molecular  Biology,  Biomedical
Center  U. of Uppsala Box 590 S-75124 Uppsala Sweden. (1947) PhD,
biophysics (london U., UK, 1973). Prof. tel. 018-174566, fax: +46
18   151759,   email:  bitnet:  alwyn@semax51  Computing,  macro-
molecules, proteins, structural biology, viruses, X-ray  diffrac-
tion


Press <RETURN> to continue, <m> to mail, <s> to save, or <p> to print:
```

Fig. 3 The selected entry is retrieved and displayed on the screen. The user may save this to a file, print it, or e-mail it elsewhere.

of indexed files which list the location of words within a collection of documents. Matching documents can then be retrieved. In this context, 'documents' may mean complete files or portions of files delimited by certain rules. Although *WAIS* can be used as a stand-alone process, it is more conveniently grafted on to a *gopher* menu for greater user friendliness. The sample IUCr *gopher* server has a *WAIS* interface allowing search of CIFs and the *World Directory*. Again, the *WAIS* program is in fact a client–server pair of processes. The information provider establishes the indexed database of words (using the *waisindex* utility), which replies to queries posed by the user's client program, or submitted *via* a *gopher* gateway. The separation into client and server programs means that we could, if necessary, modify the indexing program to modify the definition of 'words' (the byte streams within documents whose locations are actually indexed); further, the server can be modified to allow Boolean searches of multiple query terms. We are about to begin testing a modified *WAIS* server which allows **and**, **or** and **not** searches on a limited class of regular expressions (*i.e.* words *containing* the search string). This permits a greater flexibility in the queries that can be posed, without the need to construct a full relational database engine.

The third service mentioned in paper **III**, namely *WWW* or *World-Wide Web*, appears also to have potential use in the sphere of network information provision (Fig. 4). Like *gopher*, it allows documents to be retrieved from different locations; but in this case the documents retrieved are formatted, and contain hypertext links to other documents, indicated by a different typeface. One could envisage supplying an *Acta* paper in *WWW* form that could be read on-screen (and printed out on a PostScript laser printer if a permanent copy is required), with all internal references to other *Acta* papers set up as hypertext links. The reader need only click a mouse on the cross-reference to retrieve the paper referred to.

There is no difficulty with establishing immediate *gopher* and *WAIS* services: indeed, these are already available for internal use at Chester. The *gopher* server can describe files in a current filesystem, and so it is convenient to locate the *gopher* entry point at the home directory of files made available by anonymous ftp. A benefit of this approach is that the files provided by *gopher* are subject to the same level of security as any files available by anonymous ftp (that is, they should be accessible to the normal user only in accordance with the permissions set by the system administrator). The *WAIS* indexing program provides pointers to files according to their location in the filesystem, and so can be used to index the files in the anon-ftp directories as easily as files anywhere else on the system. The *gopher* server will immediately list any new files added to the current

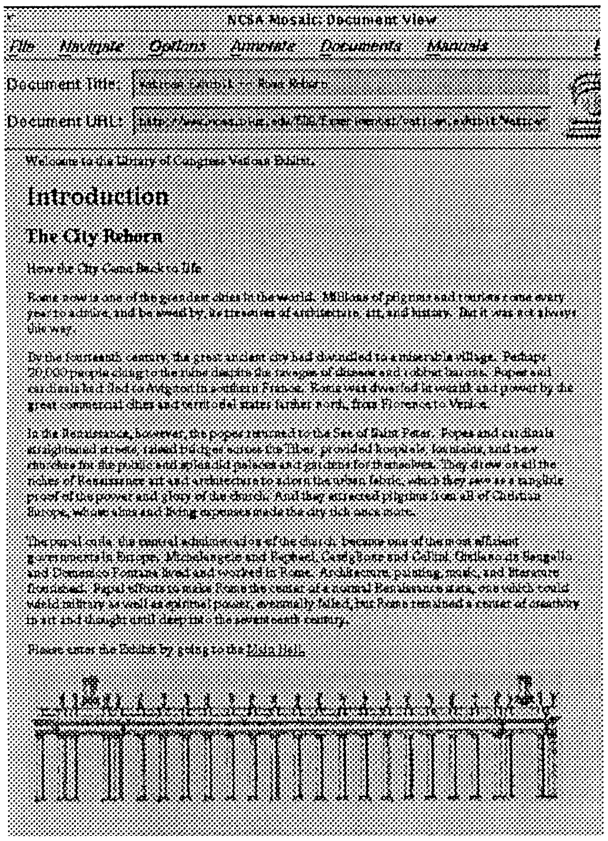

Fig. 4 An X-Windows *WWW* browser. The user has retrieved a copy from the US Library of Congress of a catalogue of Vatican art treasures. The document is attractively formatted and contains illustrations. By clicking on underlined phrases (such as *Main Hall*, at the bottom of the page), he may fetch other pages of the document, or other documents.

directory. *WAIS* indexing produces static database files, and so must be re-run when new files are added. In the case of a supposed IUCr archive and retrieval system, this could be done automatically whenever a new issue of a journal is published.

It would be more difficult to set up a fully functional *WWW* server, inasmuch as the formatting and embedding of hypertext links would need to be performed on the files to be exported. However, in the longer term this could be done, especially if the formatted documents are stored as SGML; for the internal formatting of *WWW* files is a dialect of SGML.

In summary, therefore, the provision of Internet services could proceed thus:

service	**type of file**	**timescale**
gopher	CIF, ASCII	immediate
WAIS	CIF, ASCII	immediate
WWW	CIF, ASCII	immediate
	SGML	future

These systems are already in widespread use on the Internet. They are used, for example, by the Protein Data Bank at Brookhaven and the Nucleic Acids Data Bank at Rutgers University to supply data files (in accordance with their policy of making data fully and freely accessible to the public). It is likely that the

community of users who enjoy such Internet services would come to expect similar services from the IUCr also. It is worth emphasising again that provision of these facilities uses completely standard programs, and involves the Chester staff in very little administrative overhead, at least so long as detailed control over access to the data is not required (see the next section).

Network auditing

Suppose then that we are able to supply CIFs by e-mail server, anonymous ftp, *gopher/WAIS* and *WWW*. It will be necessary to be able to monitor file traffic, and to impose a block on users exceeding their quota. There is a difficulty here, primarily with the Internet services. Traditionally, Internet protocols allow logging and authentication on a per-host basis (that is, the transactions are regarded as being between machines, rather than between individual user accounts). It is relatively easy to log transactions made to a specific host, and to block access from a specific host. But while in practice there will be many individuals who use their own workstation to access the network, equally there will be many departments or universities who access the net *via* a single mainframe or server machine.

The existing e-mail server *does* record individual account names (as it must, to mail a file to a specific account). The ftp protocol requests the user to supply his e-mail address, though there is no means of validating this. It does allow validation of the calling machine address, however. It is possible to allocate individual file-transfer accounts, each protected by a password, but the disadvantage of this is the additional administrative load it would place on the Chester end. Likewise, new versions of *gopher* are being released which permit password-based access, but it is possible that this will also burden the system administrator overmuch.

CIF subscription service

Let us assume, however, that it is possible to trace file transfers on a per-user basis, and impose a quota that the system can monitor automatically. We suggest that users who exceed their quota are barred from further file transfers, but are given the opportunity to subscribe instead to a bulk CIF supply. This would consist of a fully electronic version of *Acta* C, in the form of a set of CIFs on magnetic tape or CD-ROM, including useful software and perhaps full structure-factor sets if these are also available. Because the goods sold are individual objects, it will be possible to distribute them just as the journals are currently handled.

At the time of writing, we are about to discuss with Oxford University Press a proposal originating from them to publish the CIF archive in CD-ROM format. Our early discussions will discuss solely the technical feasibility of such a plan, for it does not yet form part of the official policy of the Union. Nevertheless, we envisage that there should be few technical problems, and we see this as potentially a valuable and effective way of raising revenue through the publication of the growing CIF archive. Later papers will report on the progress of these discussions.