CERN - DATA HANDLING DIVISION
DD/Internal Report/71/2
H. von Eicken
A. Günther
C. Piney
March 1971

Proposal for a Joint Library Mechanization and Text Processing Project

DD-cj

Chapter 1

Scope of the Project

In 1968, the CERN Library started to develop a mechanized catalogue of preprints and reports based on a weekly accessions list to which indexes by personal authors, collaboration groups, corporate authors, report numbers, conferences and subject descriptors are provided. This system uses a Flexowriter for typing the accessions list and a computer only for producing the indexes.

Since the introduction of this mechanized catalogue, the need for more satellite libraries and dissemination of bibliographic information on the extended CERN site has become apparent. National and international specialized information centres have recently started the interchange of bibliographic information on magnetic tape, a new development from which CERN may take great advantage. The time seems to have come therefore, to review the present library operation and to study possible improvement. A proposal for the development of a mechanized library system is made in chapter 2 of this report.

Since 1965, manuals for the computing facilities at CERN have been edited and typed on two Flexowriters. The paper tape of the most recent edition is kept for future updating. When preparing a new edition, the text to be retained is automatically copied from the paper tape of the previous edition, and only the alterations have to be typed manually. Although this procedure is faster than manual typing and does not require proofreading of the unchanged parts of the text, it does not provide full editing capabilities and has some other severe disadvantages inherent in the Flexowriter.

To overcome these difficulties, the Computer Users Support Group started at the end of 1969 a pilot scheme for the publishing of manuals which was implemented on the FOCUS system.

Since there is a wider interest in a more general text processing system, which could later also be applied to the composition of scientific texts, the present pilot scheme has been reviewed and a proposal for further development is made in chapter 3 of this report.

Both projects, the library mechanization and the text processing, deal with structured non-numerical data and require fully interactive data entry and editing using an extended character set.

A proposal for the joint development of the library mechanization and the text processing systems is made in chapter 4 of this report.

Chapter 2

Proposal for the Development of the Mechanized Library System

The need for library mechanization

The processing of documents in a library system involves operations at various levels. At a higher level, decisions have to be taken by specialized professional staff (e.g. selection, cataloguing and indexing of documents). At a lower level, a considerable number of clerical routines are performed (typing and dispatching orders, claiming overdues, receiving and accessioning documents, typing and sorting catalogue entries, issuing and recalling loans, etc.). At intermediate levels, some professional skill is needed for the clarification and execution of procedures such as checking of documents received, handling of unsuccessful orders.

An increase in the number of document titles received in the library affects all levels of operation while an increase in the number of users (more copies and more loans) has a predominant effect on the clerical routines. A decentralization of the library system by the introduction of branch libraries not only increases the clerical work load but also complicates the operations and calls for an effective control of the flow of documents between the branches and of the interaction between the library sub-systems.

The use of computers can greatly facilitate library operations, particularly at the clerical level, and provides a powerful tool in the control of larger or decentralized library systems.

During its first twelve years of operation, the CERN Library consisted of a central library with a large reading room and one small branch library near the south hall of the proton synchrotron. Identical card catalogues in both libraries provided the key to their document collection.

The increasing number of publications and the great number of authors of experimental papers requiring additional cards in the catalogue became a heavy burden. Furthermore, there was an increasing demand, particularly for preprints and reports, and more users had to be served on an expanding site. A working party, set up by the Library Committee in 1968, studied the users' needs and recommended that more small reference libraries should be provided, that the library catalogue must be available in these satellite libraries, and that, consequently, the library should be mechanized.

Following this recommendation, the CERN Library started to develop a mechanized catalogue for preprints and reports, based on a weekly accessions list to which cumulative indexes by personal authors, collaboration groups, corporate authors, report numbers and subject descriptors are provided. In addition, the main bibliographic data of papers presented at conferences are listed by conference in a cumulative conference index.

Present status of the mechanized catalogue

This mechanized catalogue has been in operation since the beginning of 1969 and was improved during the first fifteen months of operation to provide for smoother running. The design was based on the following concepts:

a) The system design should be simple and involve only a minimum of computer operation.

b) The weekly accessions list, needing a tight production schedule, should not be done by computer.

c) The IBM 360/30 of the Administration Department could only be used in batch operation for the production of indexes and for information retrieval.

d) The structure of bibliographic records should be compatible with other systems.

e) The system as a whole should be flexible, to make possible any future development.

Accordingly, the weekly accessions list is being typed under program control on a Flexowriter 2201, thus creating a computer-readable record on paper tape simultaneously with the hard copy. The paper tape is then converted to punched cards by an IBM 047, which was already available at CERN. The cards are read into the IBM 360/30 and the records written on disk once a week. This sequential master file is used for the production of all indexes.

Advantages and disadvantages of the present system

The present method makes it possible to print the weekly accessions list even when no computing time is available, as the print-out is done on the Flexowriter itself. This is a definite advantage over many systems depending entirely on an already overloaded computer.

Because of the restricted character set available on the Flexowriter, however, some special characters (Greek, mathematical symbols etc.) must be added by hand to the final typed list; moreover, although they are coded in the computer record they cannot be reproduced in any listing. This makes the production of bibliographies from the computer records particularly difficult.

The most serious difficulty encountered in the operation of the Flexowriter lies in the fact that control codes (tags, separators, shift codes etc.) can neither be protected nor even displayed. This makes some operations, particularly the error corrections, unnecessarily difficult, and often introduces hidden errors.

The IBM 047 in addition, has no checking facilities and therefore multiplies any errors which arise from the Flexowriter operation. Also, some characters used on the Flexowriter are not available on the IBM 047 and have to be represented by different characters, which sometimes leads to confusion and makes the correction of errors on the IBM 047 particularly cumbersome.

At its present stage of development, the current system as a whole suffers from a lack of refinement in many of the programs, inadequate authority files, susceptibility to errors and complex error-correction procedures. Although superior to a wholly manual system, it is unnecessarily complicated and time-consuming. However, in view of the serious shortcomings of the Flexowriter mentioned above, where little improvement is possible, any further development of the mechanization of library procedures should start with a complete redesign of the input system.

Concepts for the design of an input system

The existing mechanized scheme has shown that the input of structured data such as bibliographic descriptions of documents is most efficiently and safely done under program control, i.e. in an interactive mode in which a program imposes the structure and requests the operator to fill in the separate data elements.

During the input operation on the Flexowriter the program, punched on a paper tape, is read by the Flexowriter, and the instructions for the operator are typed out in a special column.

A much more efficient and flexible interactive input system could be implemented by using a CRT console on-line to a computer. As the computer can store a large variety of programs for different record structures, the transition from one program to another can be done by entering a command instead of changing the program tape on the Flexowriter. The output of instructions is also speeded up considerably on a display. This makes it possible to give instructions in natural language instead of using codes. Furthermore, data elements using a limited vocabulary (e.g. corporate authors, subject index terms, journal titles) can be composed by calling these elements out of a dictionary stored on a random access storage device. This not only saves time in the keying operation but also avoids possible typing errors. Finally, the computer can already at the input stage perform validity tests on the various data elements (length of sub-records, checking for invalid characters etc.), so that errors in the record structure can immediately be signaled to the operator by sending an error message and blocking further operation.

Such an input system can be implemented on a small computer to which one or several CRT display units are connected. This computer could also assist in the control of various housekeeping operations of the library, for instance ordering books and tracking them through the processing procedure (claiming, checking invoices, cataloguing, marking, display, ordering of additional copies), and controlling circulation.

Proposed interactive I/O system

The interactive I/O system best suited for this application should be based on a small computer with a random access mass storage (RAM) device, two CRT displays with at least full USASCII character set (upper and lower case), and a Teletype 37.

In a first stage, the system will be used for the preparation of bibliographic records for the input to a larger computer (at present the IBM 360/30 of the Administration Department).

The operator on the CRT console indicates by a command the type of record structure needed for the bibliographic description of the document to be processed. In response, the computer displays on the CRT, in the form of a kind of questionnaire, the corresponding record structure to be filled in by the operator. When a corporate author or a subject index term has to be filled in, the operator keys only the first three or four characters and pushes a search key which causes the small computer to search the RAM device containing the corresponding dictionary. All terms starting with the three or four characters keyed in are then displayed on a free part of the CRT screen, each term preceded by a number (1..., 2..., 3..., etc). The operator has only to type the corresponding number in order to complete the entry of the selected term.

The operator should not be able to over-write any tag or separator, and structural errors (e.g. omission of a data element, keying in of forbidden characters) should cause blocking of the input operation and display of an error message.

Completed bibliographic entries are stored in fixed length records on the RAM device of the small computer. At the end of the session, the new entries are printed out on the Teletype 37 for proof-reading by an assistant librarian. No checking of the spelling of corporate authors and subject index terms is necessary, as both have been taken from dictionaries in the RAM device.

After proof-reading, errors are corrected by calling the corresponding entries to the CRT consoles to edit them. In addition, the correct entries are identified as such. All records addressed in this procedure get a status code characterizing them as having been checked for correctness. Any record omitted in the procedure is signalled at the end of the week before the weekly accessions list is printed out on the Teletype and the computer records are finally transferred to the large computer to be further processed.

The Teletype 37 has 32 optional characters, in addition to the full Roman alphabet in upper and lower case, numerals and punctuation marks (USASCII set). This makes it possible to print out complete bibliographic entries including most Greek symbols. Superscripts and subscripts can be printed by using the half-line advancement under program control. Further flexibility is added by the option to print in black or red under program control.

Further development of the library mechanization

In a second stage of the project, some housekeeping routines for the library should be performed by the small computer system.

When a book or report is ordered, the major part of the cataloguing information is already available and needed for the description of the item to be ordered. This pre-cataloguing information can be updated to obtain the final catalogue entry when the document arrives. Thus, at the moment of ordering a book or report, the bibliographic description is entered, as precisely as known, via a CRT console into the RAM device, thus creating an "order file". Orders sent out to booksellers are printed on the Teletype. A computer routine checks periodically the order file and prints out claims for overdue orders.

When an ordered item arrives, the pre-cataloguing information is updated and added to the library catalogue. The entry still remaining in the order file is flagged by a status code indicating the arrival of the item. The checking of the invoice, usually coming later, is also done on a CRT terminal. Another status code indicates that the publication has been paid for. All subsequent steps in the processing of publications which need some checking or tracking are done on the basis of the record in the order file. It is, of course, very easy to order additional copies of publications for which there is a great demand.

Periodical publications (journals, annual reviews etc.) can be processed in a similar way. In this case, the checking and claiming of overdue issues are major features of the system.

In a third stage of the project, the use of the computer could be envisaged for the control of library loans.

It is very important to plan all future steps in the mechanization of the CERN Library in view of an integrated interactive system, because only such a system will offer enough flexibility and the possibility of performing a maximum of routine work by computer.

The possibility of on-line information retrieval

After full implementation of the proposed I/O system it would be desirable to establish a data link to a larger computer, preferably to the CDC 6400/7600 system, in order to transfer the corrected bibliographic records directly to a central data bank from where they can be retrieved by any interested user. Additional bibliographic information obtained on magnetic tape from other information services (e.g. INSPEC, INIS) could be added to this data bank.

The final aim of such a central information system should be defined according to users' needs.

Storage requirements for the data bank

At the end of 1970, the CERN Library held about 25,000 books and 800 periodicals. Since the beginning of the mechanization of the library catalogue for preprints and reports, over 15,000 documents of this category have been catalogued.

Assuming that all books will have to be recatalogued for the mechanized system, but that no report or preprint issued before 1.1.1969 will be added, and assuming an average record length of 600 characters per entry, the present total storage requirement for the library catalogue (books and preprints and reports) will be

2.5×10^7 characters.

Assuming further an annual growth of 10,000 items, the storage requirement for the library catalogue will be of the order of

6×10^7 characters in 1975.

An information-retrieval system including bibliographic entries selected from tapes received from other information centres (INSPEC, INIS etc.) would require considerably more storage space.

Storage requirements for the RAM device of the small computer

It is necessary to have files of all library material which is "in process" available on the RAM device of the small computer. This includes a minimum of 500 records for preprints and reports, an order file for books of 1500 records, and a file for periodicals of 1000 records, all together 3000 records.

In order to ease input and editing operations, as well as retrieval from the "in process" files, it is necessary to locate all records in fixed-length fields. It can be assumed that 1500 of these records can be placed in fields of 512 characters, while 1300 further records may need a field length of 1024 and the remaining 200 records 1536 characters, i.e. all together

2.4×10^6 characters.

The corporate author file at present contains nearly 1500 entries, to be located in fixed fields of 180 characters each, hence,

0.3×10^6 characters.

The subject index uses about 10,000 terms of up to 80 characters each; this will require

0.8×10^6 characters.

In addition, journal titles and titles of current conferences will require about

0.5×10^6 characters.

Thus, the predictable total storage requirements for the RAM device of the small computer amount to

$$4 \times 10^6 \text{ characters.}$$

Library effectiveness and staff requirements

The current system for partial mechanization of the library has been running for over two years. It provides a weekly accessions list of preprints and reports typed on a Flexowriter with computer produced cumulative indexes. Because of the difficulties of an extension of automatic operations in this system mentioned in the beginning of this chapter, all other library operations are performed manually.

An extrapolation of staff requirements for this system has to take into account that the increase of clerical effort is proportional both to the number of documents and the number of users (CERN staff and particularly fellows and visitors), heavily weighted by the introduction of satellite libraries needed for efficient service on the extended site.

Over the last ten years, during which time only the Central Library significantly affected the requirements, the allocation of one extra staff post every two years on the average has proved to be barely sufficient. It is therefore apparent that the continuation of the present system with no development of current capabilities would require at least the same staff increase. In fact, considering the increase of work load due to the recent introduction of new satellite libraries, new staff posts will be required in 1972, 1973, 1974, 1976 and 1978.

In the present system, only the Central Library and the PS Library have catalogues of all library material. A further duplication of the 100,000 and more cards in the catalogue and a maintenance of additional card catalogues in satellite libraries can hardly be envisaged. The only possibility of providing the new reading rooms with catalogues of all literature available in the library network is a complete mechanization of the library catalogue, which could then be made available on microfilm or hard copy and could also be searched on-line.

Within the field of science and technology, national and international systems for the interchange of bibliographic information on magnetic tape are developing. Examples are the International Nuclear Information System (INIS) and the INSPEC system including Physics Abstracts, Electrical Engineering Abstracts and Computer and Control Abstracts. The best use of these services can be made in a mechanized library system as suggested in this proposal. The bibliographic information available on these tapes (journal articles, reports, conference papers and books) conveniently supplements the library catalogue and therefore greatly increases its effectiveness as a source for retrieval of information on a given subject. The tapes can also be exploited for a current awareness service (selective dissemination of information). New publications to be acquired for the library can partly be selected from these records and their catalogue entries be copied from the tapes.

An integrated mechanized library system also comprises the control by computer of the circulation of library material. This not only relieves the library staff from a bulk of clerical routines but also makes it possible to keep precise loan statistics for each document. These statistics are a useful tool in the review of the acquisition policy and for the transfer of less used material from the reading room to compact deposit stacks.

As the number of satellite libraries grows, there will be an increasing flow of documents within the library network which makes manual procedures complicated but which can be most effectively controlled by computer.

The staff requirements for the conversion to, and maintenance of, an integrated mechanized library system have been studied on the grounds of experience with the current system and of the expected effort involved in the transition. It is apparent that more staff are required during the conversion to the new system, but that an efficiently mechanized library in full operation would not only offer better services but also relieve staff from clerical routines which can be performed by machines. Thus, the implementation of the cataloguing system, in which some previous experience has been gained, will free a large part of staff time needed for the implementation of the following sub-systems. On the other hand, the conversion of the present card catalogue of nearly 30,000 books into machine readable form will require a special effort.

It is hoped that the following additional staff posts will be sufficient for the implementation of the total system:

1 clerk for input operation in 1972

1 assistant librarian in 1972 and possibly 1 assistant librarian in 1973 for systems implementation

3 temporary clerks for about 18 months each for the conversion of the card catalogue.

When the new library system is in full operation, it should be possible to run it without additional staff increase until about the end of this decade except in the case of a considerable increase in demand for services (e.g. new branch libraries needing staff in attendance, information retrieval, selective dissemination of information).

In conclusion, it is evident that a continuation of the present system would run into serious difficulties, particularly because of the increasing demand for library services on the extended site. It would require increasingly more staff and become progressively more complicated from the point of view of both users and staff. On the other hand, a redesigned mechanized system developed to its full capabilities, could help to resolve many of the current and future problems and lead to a reduction of requirements for staff and an increase of effectiveness.

For this reason, it appears imperative that work should progress rapidly towards the implementation of a computer based integrated library system.

Chapter 3

Proposal for the Development of CERNTEXT (the Mechanized Pilot Scheme for Publishing of Manuals)

1. Introduction

Since 1965 the preparation and editing of all manuals for the computing facilities has been done with two electric typewriters comprising paper tape read and punch facilities. The information contained on a page of text is automatically copied to a paper tape when the typist types it. This paper tape is then used as a master whenever alterations to the text on that page are needed. It is entered into the read station of the typewriter and retyping of the page is done automatically. Since all functional commands like carriage return, line feed, backspace, etc. are also stored, the layout of the text is retained. This automatic retyping can now proceed up to the portion of text that needs alteration. While this is in progress a new paper tape is punched automatically, thus creating a new master copy. The typist then performs the modifications necessary by stopping the reading of the old tape, typing the new information and manually skipping the bad contents on the old tape. The newly entered information is automatically copied to the new paper tape. Remaining unaltered parts of the text are then typed and copied automatically. The advantage of this operation is twofold :

a) unchanged parts of a page need not be retyped manually and will be transferred free of typing mistakes. This makes proofreading less time consuming and easier, provided that changed parts are marked on the margin.

b) the automatically performed copy typing is faster than manual typing.

There are, however, several severe disadvantages in this system which are due to different reasons :

a) The typist must stay at the machine when copying is performed to be able to stop it, before the part to be changed is reached.

b) Much of the code punched on the paper tape must be known to enable the manual skipping of portions of text. One therefore needs experienced personnel otherwise the time taken to correct is more than that gained by the automatic retyping.

c) Since the functional codes are also stored all corrections causing overflow of lines or pages are very cumbersome needing deletion (skipping) and insertion of functional codes besides eventual changing of hyphenation.

d) The noise level of automatic, paper tape oriented typewriters is such, that it is difficult to perform careful and conscientious work for more than short periods of time.

e) The overall operating conditions make it tempting to the typist to modify the hard copy independently of the paper tape. It is, however essential that all corrections performed are reflected on the new paper tape and that the tapes are kept in good order in a filing system.

The experience gained over the years of operation shows that it is difficult to retain qualified personnel on work that is to quite an extent humdrum but nevertheless requires experience, carefulness and conscientiousness.

To overcome, as far as possible, the outlined difficulties and drawbacks the development of a mechanized pilot scheme for the publishing of manuals was started by the Computer Users Support Group end of 1969, beginning 1970. The external description of the scheme is given in the paper "CERNTEXT, A Mechanized Pilot Scheme for the Publishing of Manuals"

The project has now reached a state which will allow the use of all of its basic components together in the beginning of 1971. It is intended to change one manual over to the new facilities to gain operational experience and more extended testing. This manual should however be maintained in parallel on the old facilities in case of failures of the new system.

2. Present implementation of the pilot scheme

The aim in the development of the pilot scheme was to use as far as possible existing hardware and software facilities and to undertake new development only in those areas absolutely necessary and there with a minimum of effort. This led to the following way of implementation :

a) Data entry and data update is done using the FOCUS commands INPUT and EDIT. Due to the nature of these commands it is necessary to use the limited character set (i.e. FORTRAN set) available in FOCUS. To be able to represent all characters needed for the publishing of manuals an escape character has to be used (i.e. $A means A and A means a). This will certainly cause many errors, apart from the fact that it is "unusual practice" for typists.

To change this situation under FOCUS it would be necessary to design two similar commands to handle the larger character set.

b) The processing of the data files is done on the CDC 6000 series computers and the jobs are submitted via the SEND command in FOCUS. Due to the encoded form of the input data, a code conversion is necessary to ASCII character set prior to the processing. (The final output of the processing is in ASCII code to allow printing on the teletype Mod. 37).

The processing program is written mainly in FORTRAN with some assembler language routines to read and write ASCII code. Even though FORTRAN is not very suited for this type of work it was preferred to assembler language to ease development and testing. Not all of the features foreseen and necessary for production are as yet implemented. (See Appendix B of the paper about the pilot scheme for details).

c) Data output is done using a new command written for the FOCUS system to handle the ACSII character set. This command allows the use of either the teletype model 37 or the Tektronix T4002 display as output device. The command is currently being tested and should be made available for general use by March 1971.

d) To be able to use a faster printing media, a routine has been written for the IBM 1130 system which allows a card to printer operation. This routine will read binary cards produced on the CDC 6000 series computers containing the output from the text processing program. Currently it allows only the use of the existing chain/train on the IBM 1130 printer, thus only providing for upper case letters. One of the IBM 1130 printers however has got the universal character set and 10 lines/inch spacing facility to allow the use of a TN chain, comprising all characters needed. The TN chain has not yet been ordered. It would cost about 3.5K Sfr. Only a trivial modification would be necessary to the program, to use this chain.

To facilitate the use of the existing pilot scheme for gaining operational experience, the implementation of some further features into the text processing nucleus on the CDC 6000 series computers is envisaged as a short term development, these features are specifically the table of contents and keyword index facility. This together with some cleaning up will necessitate an effort of about 3 man months.

3. Proposed long term development

The pilot scheme was developed for the use of the computer documentation production. Since the SIS group in DD has expressed interest in the text processing system and would like to use it later for part of the work in the scientific typing pool, the long term proposal should cater for this. Apart from requiring more terminals for input, editing and output, this also means a further extension of the character set to include the greek alphabet and special graphics for mathematical formulae as well as larger storage requirements.

In the following paragraphs the user specifications of the proposed development are outlined and the requirements for I/O devices and storage are given.

3.1 User Specifications

A "work place" for a typist using text processing should comprise a high speed video terminal for text entry, correcting and viewing and a low speed terminal for the output of hard copies.

The system should allow her to :

a) create new documents using the layout commands described for the pilot scheme. These commands should be augmented by an optional automatic hyphenation facility and those features found missing when using the pilot scheme;

b) switch viewing of "raw data" i.e. text with embedded layout commands, to layed out text instantaneously;

c) call up from permanent storage any part of documentation in "raw data" form for viewing, correcting or processing purpose;

d) ask any part of the documentation to be printed on the hard copy device in "raw data" or processed form;

e) ask any part of the documentation to be processed and subsequently "printed" on the CØM device;

f) delete part or all of a particular document.

To fulfill all of the above specifications, it is necessary to implement the text processing system as a fully interactive system, backed up by a permanent filing scheme for "archive" data and a work storage area, into which the file to be processed is copied from the permanent filing scheme. To allow more than one user at a time, it must be implemented in reentrant code.

3.2 I/O Devices

a) Text entry, correction and viewing should be done with terminals capable of handling at least the full ASCII character set (96 graphics and 32 controls) the additional characters not contained in ASCII could be handled in encoded form, using the escape character provided for by ASCII. Since these additional symbols (greek alphabet, special mathematical symbols) do not appear very frequently the proposed solution is not a major drawback. Preferably these terminals should be displays to gain speed and to avoid noise problems. They should have their own memory with screen editing facilities under cursor control to eliminate extra software development.

b) Text output for proofreading should be done with low speed terminals capable of handling the full ASCII character set, and most of the other symbols not contained in ASCII. For this the existing Teletype Mod 37 could be used. It provides the facility of adding 32 extra symbols to the ASCII set and also allows for printing in red or black under program control.

c) Text output for offset reproduction should be done using the proposed CØM device for the main computer centre. This would allow the generation of output containing all symbols wanted. (using the graphics facility of the CØM device). As this appears to be the best solution the printer of the IBM 1130 mentioned earlier is no longer considered.

d) The number of input and output devices needed depends largely on the use made of the scheme. For the computer documentation only two devices of each type are considered to be sufficient. Besides providing two work places this is also a safety measure against breakdown of one of these devices.

3.3 Storage requirements

a) Permanent

All files containing text should be kept in a permanent filing scheme. Currently the computer documentation comprises about 2000 pages of information. Each page contains about 2100 characters, a character being represented by 8-bits. (This allows for the extended character set.) Since it is always necessary to keep two generations of each page for input to the text processing, a total storage of

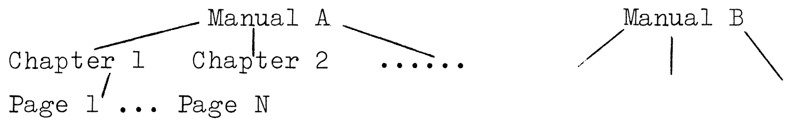
$$\approx 8.2 \times 10^6$$

8 bit characters is needed to keep the present information. Even though the use of the manufacturer's literature for the new computer complex is foreseen, it is estimated that there will undoubtedly be some CERN provided manuals besides the program library documentation.

It is at the present moment not possible to give any other estimates, specifically not for the possible use of the scheme by the SIS group. It should however be borne in mind, that not all of this information must be kept permanently on line to the computer, (i.e. private user disk packs, archive files on tape).

b) Current

The information stored in the permanent filing scheme can be seen logically separated into the following entities :



Based on the experience with the current updating of manuals and taking into account the constraints imposed by the computerization it seems reasonable to store the contents of a chapter in a file, representing the pages as logical records within the chapter. With the assumption that chapters on average contain between 15-40 pages of information and that one is working on one chapter at a time, the storage needed per user would be between

9.45×10^4 and 2.52×10^5

8 bit characters. These figures take into account the original chapter, the modified chapter and its processed output. (Original and modified chapter contain the text with embedded layout commands, while the processed output is the camera-ready layed out chapter).

4. Conclusions

The text processing program, once implemented, should not only provide writing, editorial, and publishing personnel with a system that can expedite the production of publications, but also help in saving man power and provide better working conditions. It is believed that with the Flexowriter based system about 50% of the effort is spent on other work than actual amending and correction of text.

Chapter 4

Joint Development of the Library Mechanization and the Text Processing System CERNTEXT.

A. Introduction

Proposals for the further development of two data handling applications in the DD-Division have been made in the two preceeding chapters :

Development of the Mechanized Library System
Development of CERNTEXT, the Mechanized Pilot Scheme for the Publishing of Manuals

Both proposals have many common features :

- Fully interactive data entry and editing
- Non-standard, but common character set
- Need for special 1/0 devices to handle the non-standard character set
- Textual information in both cases
- Large storage requirements for permanent data (equivalent to a few tape reels)

It is therefore advantagious to combine the two projects. The aim of this chapter is to :

- propose an implementation suitable for both applications
- specify the requirements for hardware to be acquired for the proposed implementation
- specify the requirements for software support
- suggest a procedure to be followed for the selection of the hardware
- estimate the effort for the implementation of the application programs

B. Proposed implementation

It is proposed to implement both applications on a small satellite computer to the CDC 7600 computer complex. This small computer should :

- provide the fully interactive services required for both applications, taking into account the large character set needed
- allow simultaneous access for both applications each with more than one user
- provide fast random access mass storage for work files needed by the applications during operation
- handle the communication with the CDC 7600 complex for file transfer and job initialization

The CDC 7600 host system should provide the following facilities for the project :

- permanent file store for initially about $3.5x10^7$ characters. This might grow to about 1×10^8 characters in 1975
- communication to connect the satellite computer allowing for transfer of files and job initialization for bulk processing,
- access to peripheral equipment not provided for by the satellite computer (magnetic tapes, card reader/punch and Computer Output on Microfilm device)

The implementation envisaged frees the central computer complex from two specialized interactive applications, thus avoiding excessive demands on the hardware and software of the central system. The specialized services for both applications, the fully interactive data-entry, updating and preprocessing routines, which do not require the computing power of the large system, would be developped independently on the small computer. Since bulk processing and large permanent storage facilities are provided by the host system, one can restrict the small computer to the minimum necessary for interactive operation. Only a minimum of peripherals is necessary on the satellite computer, since the main system provides back-up. This saves cost particularly for magnetic tapes and faster printing devices.

C. Hardware requirements and software support

The proposed implementation needs a small computer with backing store, video display terminals and low speed printing devices. The latter, two Teletype Mod. 37, have already been bought for the pilot schemes and can be transferred to the small computer when it is installed. Their printing mechanism needs some upgrading, namely 32 more characters and extension of the line width to 80 characters.

The remaining hardware and software requirements are specified below, subdivided into minimum requirements and further desirable features where applicable.

1.000 Small computer

1.100 Hardware

1.110 Main frame

1.111 Core memory

The size of the core memory should be initially 16K with the possibility of expansion later up to 32K, preferably in smaller blocks.

The cycle time should be around 1 μsec. Access to individual 8-bit characters is essential, which entails the requirement that a word should be an integer number of these bytes, preferably 8, 16 or 32 bits.

Further desirable features would be :

bit addressing, parity check and memory protection.

1.112 Order code and addressing modes

Apart from the standard order code for non numeric applications, the processor should provide for register to core comparisons on a byte level and automatic saving of essential registers when swapping control. Addressing in both, direct and indirect modes must be available.

Further desirable features would be :

trapped instructions, indirect and modified indirect addressing and ease of re-entrant coding.

1.113 Interrupts

Multi-level interrupts with the possibility of selective masking are essential.

1.114 Channels

The computer should be able to accomodate peripheral equipment the speed of which is widely spread. It should support computer to computer connection and provide direct memory access for the high speed peripheral equipment on a cycle stealing basis.

1.120 Peripherals

The computers must be able to connect up to four teletypes Model 37 and a number of between 2 and 8 video display terminals with industry standards interface. The transfer rate for these video display terminals could vary between TTY speed and up to 9600 bd depending on the terminal chosen.

1.121 Random access mass storage device (RAM)

The RAM device must accomodate up to 5×10^6 characters, of 8-bits each, for temporary work files and scratch store. Further storage requirements are dependent on the operating system delivered by the manufacturer. (System residence on the RAM).

Further desirable features would be :
multispindle, exchangeable disk cartridges to avoid conta mination problems.

1.122 Paper tape reader/punch

A high speed paper tape reader/punch to generate the software system and to allow for communication with the "outside world" prior to the availability of the proposed link to the CDC 7600 complex is necessary. The punching requirements for the latter might amount to 1 to 2×10^5 characters per week, in the initial stage of the project when small scale production is started.

1.123 Computer to computer link

The small computer should allow for block transfer of information to and from the CDC 7600 complex. The speed of the transfer could be anything between 4800 bd and 48Kb. Direct memory access for the interface is desirable.

Since at the current point in time no definite plans for the type of connection have emerged, no further details can be specified. The characteristics are to a large extend dependent on the way INTERCOM in the CDC 6000 front end computer(s) converses with its peripheral devices.

1.130 Installation requirements

The operating conditions should allow the computer to be installed in a location containing neither special air-conditioning equipment nor specially regulated mains supply. In general the conditions would correspond approximately to those in ordinary offices, i.e. a 50 Hz mains supply having a voltage variation of ± 15 % and a frequency stability of ± 2 %.

1.140 Maintenance

Remedial maintenance should be available from manufacturers personnel at a few hours notice and adequate diagnostic programs should be abailable to CERN personnel to establish the quality of machine performance at any time.

1.200 Software support

1.210 Operating system

The manufacturer's operating system should provide time sharing for several users and a variety of tasks or at least contain 'hooks' to include this facility. It should support random access mass storage devices for user work files and scratch store. Input/output operations should be overlapped with other processing. The input/output system should be sufficiently flexible to permit straight forward addition of software for new devices, not necessarily contained in the manufacturer's range. Documentation should be exact and complete.

Further desirable features would be :

The operating system should be disk oriented.
Its core residence should be minimal and adaptable to customer needs. The allocation of resources, core and disk buffers, should be dynamical.

1.220 System modules

The following system modules are required :

1.221 A symbolic assembler, preferably with macro possibilities, producing relocatable binary code.

1.222 A linkage editor for relocating and linking groups of modules produced by the assembler and/or contained in the library.

1.223 File handling facilities for random access and sequential files, including handling of directories.

1.224 An editor, preferably disk oriented, for creation, deletion and updating of program and data files.

1.225 Utility routines to transfer files between peripheral devices.

1.226 Debugging aids in form of break pointing, selective dumping and memory modification.

A further desirable feature would be :

1.227 The input/output routines delivered with the system should be able to handle the full ASCII characters set.

2.000 Video display terminals

Initially two video display terminals should be connected to the small computer system to allow the development of the application packages. This number will have to be increased, once production has started for both applications. A current estimate is a total of 4 - 5 terminals. If at a later stage more users are added for the text processing system and further stages of the library mechanization are implemented, the required total may rise to 8 terminals.

2.100 Device requirements

- Industry standards interface
- Transfer rate of 1200 to 2400 bd
- Input and output of the entire ASCII set of 128 characters (96 graphics and 32 controls)
- Displayable screen area of at least 20 lines, 80 characters each
- Current screen position marked by a computer positionable cursor
- Screen readable in a normally illuminated room
- Full or half duplex transmission modes to improve the man/machine interface.

Further desirable features would be :

- Extension of the character set to ncn standard symbols.

- Integral display memory larger than screen size, to allow 'windowing' through locally stored text.

- Editing the integral display memory directly from the console, without computer intervention.

- Computer setting up of formats, for field structured data, on the screen, allowing purely local control of structuring.

D. Computer selection procedure

The Small Computer File presents a collection of small computers generally tendered by manufactures to CERN and may therefore be used as a basis for selection. The following principles are to be applied :

- The small computer must fulfil the stated minimum hardware and software requirements.

- In case of equally prices offers, preference should be given to those that satisfy best the further desirable features.

- Preference should be given in general to those computers tendered by European manufacturers.

- Only those non European manufacturers should be considered who provide maintenance and part service from subsidiaries in Switzerland or the area of Geneva.

- The price of the initial configuration comprising :

  information processor with console
  16 K core memory
  5 x 10^6 characters RAM device
  paper tape reader/punch
  interfaces for two video display terminals
  interfaces for two Teletype Mod. 37

  should not be significantly more than the budget forecasted for 1971.

- The further extensions of the small computer with respect to

  additional core memory
  additional RAM device capacity
  additional interfaces
  the computer to computer link

  in the following two years should fit within the budgetory limitations in force at that time.

E. Implementation effort needed

The effort needed to implement both applications depends, at least in the initial stages, largely upon the software provided by the manufacturer. The implementation should be based on the manufacturer's standard operating system even if some efficiency is lost. This policy should allow the use of later versions of the software provided by the manufacturer. The main priority of the implementation should rather be to provide early production facilities for the library and computer documentation services.

To achieve this it is necessary to divide both applications into logically independent stages, each of which adds to the production facilities provided so far. Some constraints are imposed on the time scale by the time table for the installation of the 7600 computer.

1.0 Effort needed for library mechanization

Since the beginning of 1969 the CERN Library has operated a mechanized catalogue for preprints and reports. For the future a new record structure is needed in order to cover not only preprints and reports, but also books and periodicals. International standards introduced since must be accounted for. Due to these changes the processing programs currently used on the IBM /360-30 at CERN are no longer adequate.

After consideration of the above reasoning and the specifications laid out in chapter 1 of this paper the following stages are proposed for the development of the library mechanization.

1.1 Design and implementation of the fully interactive data entry and correction routines for library data on the satellite computer. From the outset, look-up and selection facilities from dictionaries containing reoccuring index terms and corporate authors should be provided. The records containing the weekly accessions should be compressed before they are transmitted to the host computer. (8 man/month).

1.2 Adaptation of the processing routines to the CDC 7600 which :

- provide sorts against various keywords
- update a cumulative author index
- set up a KWIC index
- list sorted data in desired formats
- update a master file

Furthermore conversion routines to adapt the accumulated IBM /360 - 30 data to the new format and to prepare the various output listings in ASCII code for the precessing by the COM device are needed. (12 man/month).

1.3 Design and implementation of routines on the satellite computer to allow :

- precataloguing of books etc, at the time of order
- claiming of overdues
- checking of invoices
- checking of periodicals,

(6 man/month)

2.0 Mechanization of Computer Documentation (CERNTEXT)

The Computer Documentation Office in the past, and since mid 1970 the DD Secretariat, produce the computer manuals which are published by CERN using paper tape oriented typewriters. As a first step towards mechanization a pilot scheme for text precessing is now introduced using facilities provided on the current central computer installation. No operational experience has as yet been gained. Some of the essential features are not yet implemented.

Considering this and taking into account the specifications for the further development of CERNTEXT, the following stages are proposed :

2.1 Completion of the pilot scheme, cleaning up and providing automatic index creation for table of contents and keywords.

(3 man/month)

2.2 Use of the pilot scheme by the DD Secretariat to gain operational experience. Initially, one manual should be transferred to the new scheme, but maintained in parallel by the old scheme. After about 3 - 4 months, depending on success, further manuals should be transferred. Maintenance of the pilot scheme should carry on, but major modifications should not be envisaged for it.

2.3 Transfer of the pilot scheme to the satellite computer according to the proposals made in chapter 2. At this stage the modifications deemed necessary from the operation of the pilot scheme should be implemented.

(12 man/month)

2.4 Augmentation of the text processing programs by adding automatic hyphenation facilities. At the same time automatic spelling checks might be considered.

(4 man/month)

General system effort

So far only details about the effort needed to implement the first stages of the proposed applications have been given. For an overall impression however is it necessary to specify the effort needed to acquire knowledge of the hardware and software of the satellite computer, of drivers for non-standard equipment (Teletype Model 37, video display terminals), of the interface between the application programs themselves and the system, and of the interface between the satellite computer and the CDC 7600 complex. It is rather difficult to give estimates for these for the following reasons :

- the computer has nct yet been selected
- documentation and listings describing the software system in sufficient detail are not available. (Some of the systems are still being checked out).
- our own plans with regard to the CDC 7600 are still in their early stages.

Tentatively one can quote the following :

3.1 Up to 6 man/month are needed to study the hardware and software.

3.2 It should be possible to design and implement the necessary drivers within 4 man/month depending upon the amount of sophistication needed.

3.3 The general design of interfaces between the application programs and the system will probably take up to 3 man/month.

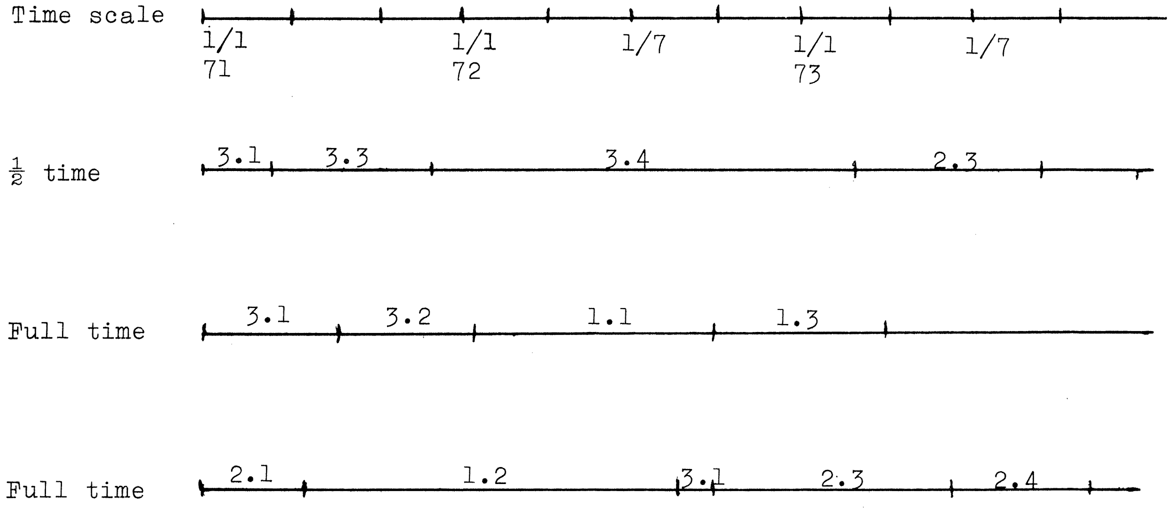
3.4 Design and implementation of the communications package with the CDC 7600 complex including some routines on the CDC 7600 side can take up to 6 man/month effort.

Summary

The total effort to implement the proposed applications on a satellite computer of the CDC 7600 complex can be summarized as follows :

1.0	Library mechanization	
1.1	Interactive data entry etc	8 man/month
1.2	Adaptation of processing routines	12 man/month
1.3	"house keeping"	6 man/month
2.0	Computer Documentation	
2.1	Completion of pilot scheme	3 man/month
2.2	Use of pilot scheme	-
2.3	Transfer to satellite computer	12 man/month
2.4	Hyphenation	4 man/month
3.0	General system effort (tentatively)	
3.1	Understanding of hardware and software	6 man/month
3.2	Design and implementation of drivers	4 man/month
3.3	Design of interfaces	3 man/month
3.4	Communication with CDC 7600	6 man/month
	Total effort involved	64 man/month

If one assumes, that two programmers are devoted full time and one programmer half time to the implementation of the project, the following time scale and distribution of work could be envisaged :



This time scale would allow the CERN Library to transfer all of its current activities from the IBM /360-30 and flexowriter to the satellite computer by August 1972. In February 1973 the "house keeping" routines should be ready, thus releaving one full time programmer, who at that point in time could begin the study of an information retrieval system for the library.

The transfer from the pilot scheme for computer Documentation processing will be delayed most since the new scheme will not be available before mid 1973.

The connection to the CDC 7600 complex is expected for the third quarter 1972. This is consistent with the current plans for the installation of the 7600.

In the above time scale it is furthermore assumed, that the small computer is delivered to CERN not later than August 1971.