Preparation of Personal Bibliographies Using a Large Computing Facility

F. R. LIPSETT* and A. B. HAYCOCK
National Research Council, Ottawa, Ontario K1A 0R6, Canada

Received June 5, 1981

Large modern computing facilities, which normally include text editing, sorting, and searching, are now available to most scientists. Such features are invaluable for a personal bibliography, which may be prepared according to the taste of the user. Our method requires only that the references be less than or equal to 120 or 240 characters (spaces) in length. The references may be written randomly, in which case searching is performed for retrieval, with some order, when searching and sorting may be used for retrieval, or with extensive order, when, in addition, indexes may be prepared by a single command. Two bibliographies with extensive order are described. Suggestions for starting a bibliography by those unfamiliar with computers are given in the Appendix.

INTRODUCTION

Few papers on the use of computers for personal bibliographies have been published. Lipsett and Blair[1] described a bibliography recorded on magnetic tape. It used specially printed cards for preparing references for entry into the computer, as well as special programs, and was open ended with respect to length. It was designed for producing a large annual bibliography for publication[2] but was also used as a personal bibliography. It is no longer employed by us because the systems described below are simpler to use and are adequate for our changed interests. Calvin[3] described a personal bibliography which used coded entries and special programs on a small computer. Reprint requests and 3 × 5 file cards were automatically printed. No subject classifications were included, so retrieval was entirely by searching. Van Ree[4] described an ingenious system using a programmable calculator with a cassette drive. A maximum of 900 references could be stored, but each was at most only 40 characters, and special coding was required to reduce keywords to 5 characters. Retrieval was by searching. More recently Kaetzel, Glass, and Smith[5] described a system for indexing references using a minicomputer. They used a fixed format with a maximum of 372 characters per reference and fixed fields, for various lengths, for 14 identifiers such as author 1, author 2, title, or keyword 1. Programs specially written in FORTRAN V Level I were employed.

Computing facilities have improved since most of the foregoing papers were published. Most scientists now have access to a large computing facility. In our case we use the facilities of the National Research Council Computation Centre (NRC CC). Big computing facilities have advantages in that powerful text editing, sorting, and searching facilities are normally available. They are also able to handle and store large amounts of data rapidly and cheaply. Many have commercially developed data base management systems such as STAIRS, TOTAL, or EASYTRIEVE which could be used for a bibliography. These may in fact be overly powerful for a computer application such as that discussed here. Modern computers offer a host of support facilities and are therefore well suited for use for a personal bibliography, allowing great leeway in the way a bibliography may be prepared and used. We will describe how facilities at the NRC CC were used to prepare two bibliographies and will indicate how they might be used to produce bibliographies in different styles.

DESCRIPTION

Requirements of a Personal Scientific Bibliography. A personal scientific bibliography will normally be used for publication of specialized bibliographies, for relocating references in a library, and for relocating reprints or similar material in one's own files. If bibliographies are to be published, the record length for each reference must be very large or open ended,[1] and the retrieval of references may require cumbersome procedures. If, however, the bibliography is used mainly for relocating material, the references need not be open ended but may be limited to some convenient length.

The number of characters needed for a reference depends on the nature of the material and on the system used for entry and retrieval. In a small sample of references taken from the authors' early bibliography[1] the number of characters for an author was 6–21 and for a title 25–119, for a total of 31–140. Most references had more than one author. In a sample of 16 books listed in the "Books Received" column of a recent issue of *Science* (*Washington*, *DC*), the average number of characters per book was 167, although the description of one book required 318 characters. Thus about 150–200 characters per reference would seem a good number to aim for, although fewer would often be adequate with appropriate abbreviation.

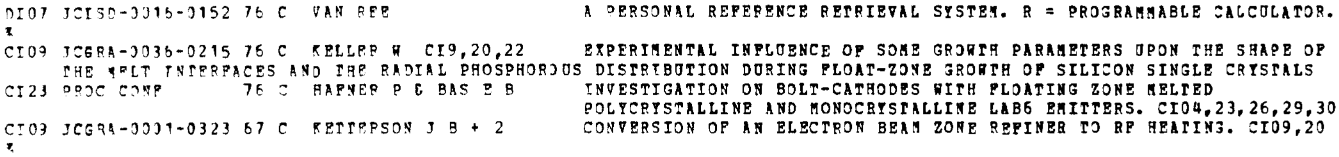
The number of characters per line available at most computer terminals is usually between 80 and 150, with printers normally limited to 133 characters per line. Thus 120 is a reasonable number of characters to choose for a line length but is too few for many references. However, it is possible, as we shall see below, to combine two lines to make a single reference of 240 characters. Therefore bibliographies including references of up to 120 or up to 240 characters may readily be prepared.

Provided each reference is of appropriate length, several arrangements of the bibliography are possible.

(1) Random: The user may enter references in any manner, for example, sometimes starting with an important author, sometimes with a date, or sometimes with a keyword. Some references may contain more information than others, and so on. References from such a bibliography would be found by searching for keywords.

(2) With some order: The user may, for example, start each reference with the first author's surname or with a subject classification. Or he might use the final columns of the reference for the date or some other identifier. It would be possible to sort the bibliography by author or subject, if they were always in the same columns, and also to search it by keyword as before. Such a system might be much better than the first with little more effort.

0095-2338/82/1622-0001$01.25/0



```
DI07 JCISD-0016-0152 76 C  VAN REE                    A PERSONAL REFERENCE RETRIEVAL SYSTEM. R = PROGRAMMABLE CALCULATOR.
%
CI09 JCGRA-0036-0215 76 C  KELLER W  CI9,20,22        EXPERIMENTAL INFLUENCE OF SOME GROWTH PARAMETERS UPON THE SHAPE OF
     THE MELT INTERFACES AND THE RADIAL PHOSPHOROUS DISTRIBUTION DURING FLOAT-ZONE GROWTH OF SILICON SINGLE CRYSTALS
CI23 PROC CONF        76 C  HAFNER P & BAS E B        INVESTIGATION ON BOLT-CATHODES WITH FLOATING ZONE MELTED
                                                      POLYCRYSTALLINE AND MONOCRYSTALLINE LAB6 EMITTERS. CI04,23,26,29,30
CI09 JCGRA-0001-0323 67 C  KETTERSON J B + 2         CONVERSION OF AN ELECTRON BEAM ZONE REFINER TO RF HEATING. CI09,20
%
```

Figure 1. Examples of references in the data set REFEX2.

Table I. Format of 240 Character References

columns	contents
1-2	Format and main subject. Headings use AA-BZ, ZA-ZZ, and additional lines in the sub-subjects. Main subjects include crystal growth and related subjects.
3-4	Secondary subjects. Up to 99 permitted within each main subject.
5	Blank.
6-20	Journal coden or identifier such as PROC CONF, THESIS, BOOK, etc. If journal columns 6-10 give the American Chemical Society coden for the journal, columns 12-15 give the volume, and columns 17-20 give the first page.
22-23	Year (last two digits).
24	Blank.
25-27	Coded remarks. A for copy of abstract or title page on file; C for copy on file; F for language not English.
28-51	Author(s), starting with surname of first author. Sorting by author is currently done on the first six columns (28-33), so what comes after matters little. Author headings start with AAAAAA and end with ZZZZAA in columns 28-33.
52-120	Title and remarks. Remarks follow the title and are prefaced by "R=". If the reference has more than one subject heading it is normally noted here.
2nd line: 1-120	Continuation of title and remarks if required or "%". If this line is not needed, the "%" will cause it to be erased in the author and subject indexes, thus eliminating unwanted blank lines.

(3) With extensive order: Identifiers such as the subject, journal name or coden, first author's surname, and so on are always placed in the same columns. The references may then be sorted on these columns to give, for example, subject and author indexes. Searching by keyword is also possible. The author's bibliographies are prepared in this way. Author and subject indexes may be prepared, from a list of references arranged in random order, by giving a single command at the computer terminal.

Facilities of the NRC CC. The NRC CC operates under the IBM Time Sharing system (TSS) on an IBM 3033-N8, supporting both conversational time sharing and batch computing. There are about 200 commands within TSS, and from it the user may enter other computing subsystems. These include facilities for text editing, a number of statistics, scientific calculations, data manipulation, simulation, plotting, and so on. A large library inquiry system, CAN/OLE, with a number of online search services and data bases including several million references, is offered by the Canada Institute for Scientific and Technical Information (CISTI) and resides on the NRC CC computer. Users are connected from all parts of Canada through the Bell DATAPAC network.

Of the facilities offered by NRC CC only TSS and EDITOR, a text-editing system, are required for the bibliographies. The essential features are as follows:

(1) The ability to handle lines of 240 characters: Such lines never appear before the user, who prepares references as pairs of lines of 120 characters. The pairs are combined into lines of 240 characters in EDITOR, sorted in TSS in this form, and split back into pairs in EDITOR.

(2) The ability to sort: The TSS command SORTDS will sort a data set (or set of references) alphabetically or numerically or both, one or more times. For an author index only an alphabetical sort is required. For the subject index, however, a sort with three "keys" is necessary. This has the effect of an alphabetical sort for the main subject, a numerical sort for the secondary subject, and an alphabetical sort for the author.

(3) The ability to group sequences of commands called procedure definitions (PROCDEFS): In effect these give the user the ability to create his own TSS commands. It is not necessary to learn a programming language. The bibliography preparation and search commands are implemented by means of PROCDEFS.

(4) The ability to search the references using a keyword: This is simply done by using a PROCDEF.

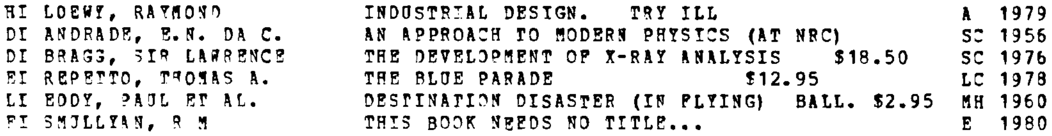
Description of bibliographies. (a) *Outline.* References are entered (typed) online into a workspace at a computer terminal by using the EDITOR system. When the entries are completed the workspace is saved as a data set on direct access storage. It may subsequently be corrected, added to, and used as required. The order in which the references are entered is immaterial, but they must have a certain format in order to allow future operations. Certain columns are reserved for sorting, and others are always filled in a certain order for neatness. Data sets consisting of headings for the author index, subject index, and results of keyword searches are also prepared. One command is used to combine the references and headings into an author index with all references listed alphabetically by author and a subject index in which references are listed by main subject heading, secondary subject heading, and alphabetically by author within each secondary heading. The indexes are prepared at intervals and suffice for most purposes. However, another command can be used to search for references containing a certain keyword. If any are found they are sorted into lists by author and subject.

(*b*) *Long Form of Bibliography* (*240 Characters per Reference*). Several examples of references as they appear in the basic data set, here named REFEX2, are shown in Figure 1 and an explanation is given in Table I. Tabs are used for aligning the coden, author, and title in Figure 1. The first reference, by Van Ree, is in main heading DI and secondary heading 07. The reference is in this journal, volume 16, page 152 (1976). A copy is on file. We forgot to put in the author's initial, but this does not effect sorting since only the first six letters are used. A remark on the use of a programmable calculator gives an indication of the contents of the paper. One line was adequate for this reference, so the second line consists only of "%" and does not appear in the indexes. The next reference, by W. Keller, is in the *Journal of Crystal Growth.* Its title is unusually long and takes almost the whole second line. Because it was so long a note of the subject categories was made in the space normally used for authors.

Frequently a reference falls into more than one subject category. In this bibliography the reference must be entered into REFEX2 once for each subject heading. This could be

Table II. Data Sets for the Long Form of Bibliography

type of data set	name	description
input	REFEX2	List of references. See Figure 1 and Table I.
input	SUBHD4	Headings for the subject index. These use AA, AB, AC,. . .CA01, CB01, CC01,. . .on the left for headings at the beginning of the subject index and at the beginning of each subject. At the end ZA, ZB, ZC,. . .are used for notes.
input	AUTHH3	Headings for the author index. For the beginning AAAAAA, AAAAAB,. . .are used in columns 28–33. Similarly ZZZZAA, ZZZZAB,. . .are used at the end for notes.
input	SEARCH·AUTHHDS	Headings for the results of keyword searches listed by author.
input	SEARCH·SUBHDS	Headings for the results of keyword searches listed by subject.
output	AUTHOR·INDEX(0)	Current version of the author index, produced by BIBLIOG operating on REFEX2 and AUTHH3.
output	SUBJECT·INDEX(0)	Current version of the subject index, produced by BIBLIOG operating on REFEX2 and SUBHD4.
output	KEYWORD·SEARCH(0)	Latest result of a keyword search, produced by SEARCH1 operating on REFEX2, SEARCH·AUTHHDS, and SEARCH·SUBHDS.
ref	COD9	A list of codens, for use when entering references into REFEX2.
ref	SHD1	A list of subject headings (a subset of SUBHD4) for use when entering references into REFEX2.



```
HI LOEWY, RAYMOND          INDUSTRIAL DESIGN.   TRY ILL                  A  1979
DI ANDRADE, E.N. DA C.     AN APPROACH TO MODERN PHYSICS (AT NRC)        SC 1956
DI BRAGG, SIR LAWRENCE     THE DEVELOPMENT OF X-RAY ANALYSIS    $18.50   SC 1976
EI REPETTO, THOMAS A.      THE BLUE PARADE                  $12.95       LC 1978
LI EDDY, PAUL ET AL.       DESTINATION DISASTER (IN FLYING)  BALL. $2.95 MH 1960
FI SMULLYAN, R M           THIS BOOK NEEDS NO TITLE...                   E  1980
```

Figure 2. Examples of references in the data set BOOKADD.

done by typing the reference as many times as necessary, but this would be boring and subject to error. Therefore a command called NEWSUBJ was written which copies a reference, changing only the subject categories as required, for up to seven categories. Thus after entering the Keller reference the following was typed:

NEW-SUBJ CI09,CI20,CI22

The reference was automatically reentered with CI20 and CI22 as subjects. The reference appears once in each category in the subject index, but three times in the author index.

In addition to REFEX2 several other data sets are required. Some are produced by the commands, while other are kept for reference. They are listed in Table II. The subject headings (SUBHD4) include spaces, main headings, subheadings, underlinings, and notes at the end. At the time of writing there were three main headings and a total of 78 subheadings. These are added to from time to time. Headings for the author index appear only at the beginning, with notes at the end. The headings for keyword searches appear only at the beginning, and subject headings are not included. Data sets with the suffix "(0)" are members of a generation data group in which the number in parentheses is the relative generation number. The latest version of the data set is suffixed (0), the one before that (–1), before that (–2), and so on. The designer can specify how many generations are maintained. When this limit is reached, the oldest generation is erased as each new generation is created. Each time BIBLIOG is used the generation advances. Only the latest generation of the author and subject indexes is retained. However, 10 generations of KEYWORD·SEARCH are kept, since these may remain of interest.

Three commands are used. NEWSUBJ was described above. The command BIBLIOG produces author and subject indexes in the following manner (with many details omitted). REFEX2 and AUTHH3 are combined in the EDITOR system. A "$" sign is placed at the end of every odd line, and each pair of lines is combined to form lines of 241 characters. (EDITOR can accomodate up to 256 characters per line.) The resulting data set is sorted alphabetically on columns 28–33. The sorted data set is moved back to EDITOR, where it is split back into pairs of lines at the "$" signs, and the "$" signs are then erased. The resulting author index is printed with the name AUTHOR·INDEX(0). The subject index is prepared similarly, using the headings from SUBHD4, but three sets of keys are used. The first key is alphabetical in columns 1 and 2 and places all references and headings in main headings. The next key is numerical in columns 3 and 4. It arranges the references by secondary heading within each main heading. The final key is alphabetical in columns 28–33 and arranges the references alphabetically by author within each secondary subject. The resulting data set is split back into two line pairs as above and printed as SUBJECT·INDEX(0). Although the command goes through many stages, all the user has to do is type "BIBLIOG" at his terminal, and the indexes will be prepared, printed, and mailed to him.

The third command is called SEARCH1. This searches REFEX2 for references containing a keyword. The line pairs are combined as in BIBLIOG and searched (new) line by (new) line for the keyword. If the keyword is not found in a line, then that line is erased. Thus a new data set consisting of references containing the keyword is left. If REFEX2 is not available, the search does not proceed, and if there are no hits, the user is informed. Otherwise the references are sorted, as in BIBLOG, into author and subject classes. The headings are simple, however, and the indexes are printed as a single output. The user merely types "SEARCH1 KEYWORD", where the keyword is chosen as desired.

(*c*) *Short Form of Bibliography* (*120 Characters per Reference*). We also maintain a bibliography of books of scientific, historical, and general interest. This and the long form of bibliography were prepared at roughly the same time. Some examples of references from the basic data set, BOOKADD, are given in Figure 2. The letters in columns 1 and 2 are for headings and main subjects in the subject index. The author(s) are listed in columns 4–28, and the title and remarks in columns 29–75. Columns 76–78 give the subject or secondary subject and columns 78–82 the year of publication. In columns 1 and 2 HI denotes biography and autobiography, DI science, LI modern history, EI law and crime, and FI essays, philosophy, and miscellaneous. In columns 76–78 A denotes autobiography, E denotes essays, and the remainder are equivalent to the main subjects. These columns are partly redundant and could be moved to the left, similar to the long form of bibliography, but have been left in place partly from laziness and partly because of their alliteration. In the Loewy reference "TRY ILL" means try an interlibrary loan, as at the time of writing the book was not held by NRC. Only 82 characters of the 120 available are used, but this has proved adequate. Originally the bibliography was prepared for card format with 80 characters.

This bibliography assembly includes three input data sets: BOOKADD, the basic list of references, BOOKHD, which includes headings and notes for the subject index, and WRHD, which includes headings and notes for the author index. Output data sets include a subject index called BOOKLIST(0), an author index called AUTHLIST(0), and a keyword search output called RESULTS·SEARCH(0). These are generation groups, and as before, only the current generation of the first two is retained, but 10 generations of the third are retained. A command called BOOKS is used to prepare author and subject indexes. It is similar to BIBLIOG but much simpler since there is no need to catenate and subsequently split the references. Sorting for the author index is on columns 4–9. For the subject index only two alphabetical keys are necessary: the first in columns 1 and 2 for main subjects and headings and the second in columns 4–9 for putting the references in alphabetical order by author within each main subject. No key is required for secondary subjects, which are shown in columns 76–78.

For searching this bibliography a command called SEARCH is used. SEARCH was written in a general form so that any data set of up to 120 characters per entry can be searched. The user gives the command "SEARCH DSNAME,KEYWORD" where DSNAME is the name of the data set to be searched (normally BOOKADD) and KEYWORD is the keyword for which to search. The operation of SEARCH is similar to that of SEARCH1. If the data set specified cannot be found, or if there are no hits, the search is terminated. If there are hits, they are listed in the order found (but not sorted), printed, and delivered to the user.

(*d*) *Other Possible Versions of* The *Bibliographies*. It is easy to imagine different versions of the bibliographies. A simpler version of the short form might delete one or more of the main subjects, secondary subjects, year, or alignment of the titles. If all these were deleted, the bibliography could still be searched, and if the author's surname was always on the left, it could be sorted alphabetically by author. Another simple version might include only authors(s), title, and a keyword where the title lacked an appropriate word for searching.

More complicated versions may also be visualized. For example, upper and lower case could be used. In the long form a routine for eliminating, in the author index, duplicate references prepared by NEWSUBJ for the subject index would be an improvement. The author index might also be improved by sorting each author's papers by year.

More elaborate search procedures using Boolean logic (and, or, etc.) could also be used. Such procedures are routinely used for searching large data bases and are available at the NRC CC[6] but have not yet been found necessary. The reader may thus devise a bibliography to suit his own taste.

RESULTS AND AVAILABILITY

The bibliographies have been in use for about 2 years and have proven satisfactory. Several alterations in details were made following the original formulation, but the forms shown have served well. Entering a reference in the long form takes a modestly skilled typist about 2 min, including multiple subjects. Updates are made at intervals of about 2–6 months. The author and subject indexes meet most requirements for retrieving references, and searches have seldom been made. Copies of papers are filed according to main and secondary subjects. A copy can normally be found within 1 min if it appears in an index and is correctly filed.

It should be easy to start a bibliography similar to the one described here at any large computation center. A listing of all commands described here, including comments, will be supplied on request. It may be possible to adapt them to other systems with few alterations.

APPENDIX: SUGGESTIONS FOR THOSE UNFAMILIAR WITH COMPUTERS

You should first decide the form you would like your bibliography to take. The bibliographies described above can serve as models, and we will send copies of them, as well as the headings, on request. It is important to include only essential material and to arrange the material to facilitate sorting and retrieval. Next you should enlist the aid of one or more members of your computing facility and get some instruction in the use of the computer. This might take the form of a short course or a session at a computer terminal. It is essential to have some personal instruction and to have someone to turn to when the inevitable mistakes are made while first using the computer. There are psychological difficulties—sometimes outright fear—in learning to use a computer, and personal assistance at the outset is necessary for surmounting them. Your teacher should also be able to adapt the PROCDEFS (which are simply ordered groups of computer system commands) described above for your computing facility, should they be wanted. If you are suitably placed you may be able to persuade a secretary or technician to go through this initiation or to enlist the help of one who has. You do not have to learn a programming language.

You will normaly work online, that is seated at a computer terminal. The terminal may be a hard-copy type similar to a typewriter or a video type (CRT or VDT) similar to a television set. We usually use a hard-copy type because their line length is normally greater than for video types. Your first job will be to set up your references, headings, and PROCDEFS as required for your bibliography. The bibliography may, after production, be written out at the terminal, but this is not normally done, especially if the bibliography is large, since the terminal is slow and costly in comparison with a line printer. This device is a standard item in any large computing facility and prints material at several hundred lines per minute without tying up the computer. The line printer is normally located at the computing center while the terminal may be distant, so a delay is incurred in transporting the printed bibliography to the user. Excerpts from the bibliography and the results of searches may of course be printed at the terminal. Figures 1 and 2 are photographs of the work of a line printer.

It is technically possible to utilize a distant computing facility, for example, NRC CC, but not advisable because of the need for personal interaction and for transmitting printouts. The hardware and software requirements are not extensive and will likely be available on any modestly configured computer system. Such a system should have basic terminal communication support, enough direct-access storage to hold your bibliography, and some off-line device on which to print your results. The entire bibliography system is written by using operating system commands and does not make use of any programming languages. It does, however, require the use of sorting and key editing facilities.

REFERENCES AND NOTES

(1) Lipsett, F. R.; Blair, F. D. "Bibliography Preparation by Computer". *J. Chem. Doc.* **1968**, *8*, 26–29.

(2) Lipsett, F. R. "Energy Transfer in Polyacene Solid Solutions VIII: A Bibliography for 1968". *Mol. Cryst. Liq. Cryst.* **1969**, *6*, 175–204.

(3) Calvin, W. H.; "A Computer-Assisted Personal Literature Reference System". *Comput. Programs Biomed.* **1972**, *2*, 291–296.

(4) Van Ree, T.; "A Personal Reference Retrieval System". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 152–153.

(5) Kaetzel, L. J.; Glass, R. A.; Smith, G. R.; "A Computer Data Base for Indexing Research Papers". *NBS Tech. Note (U.S.)* **1980**, *No. 1123*, 90.

(6) Green, R. A.; "Draft Manual for the NRC Information System"; National Research Council: Ottawa, Canada, 1981.