volume 11
number 2

ISSN 0022-2240

journal of library automation

june, 1978

jlauay 11 (2) 89–184 (1978); j libr automat

The **CLSI NEWSLETTER** is published quarterly to keep you informed about the latest developments in library automation. Articles include in-depth feature stories and case histories covering important subjects, such as those shown above.

The **CLSI NEWSLETTER** explains how your colleagues are receiving the numerous benefits which can be obtained through the automation of library functions and procedures. CLSI Users' Groups activities are covered regularly to highlight the many ways that libraries are sharing information and resources in cooperative projects.

In short, reading the **CLSI NEWSLETTER** is one of the easiest ways to keep up with the rapidly changing field of library automation. And, it's available to library professionals at no cost, as a service of CLSI.

Just fill out the coupon below, and we'll add your name to the thousands of librarians and systems professionals who now receive the **CLSI NEWSLETTER**.

Name _____

Position _____

Library _____

Address _____

_____

City, State, Zip _____

CLSI

The Leader in Library Automation

C L Systems, Inc. • 81 Norwood Ave. • Newtonville, MA 02160 • (617) 965-6310

JOURNAL OF LIBRARY AUTOMATION

Volume 11, Number 2: June 1978

CONTENTS

JLAUAY 11 (2) 89–184 (1978); J Libr Automat

Cast your vote for...

Quality Data Base Services—At SDC we are progressive, using our experience to bring you better information retrieval. Now, you will find a refined ORBIT® (the fourth in a series of planned developmental efforts) and data bases moving to standard on-line format.

Innovative Information Retrieval Tools—Order original documents on-line, tailor reference printouts to meet your specific search needs, and save time and money with the system sophistication found only in ORBIT.

Accountability in Product Delivery—a variety of channels is available for user support and feedback, including access to our technical staff to discuss and *respond* to your information needs, and free computer time on new data bases.

Quality Services, Innovative Tools, and Accountability are what you vote for when you call SDC Search Service's ORBIT Retrieval System. Your terminal and telephone are truly your on-line polling place.

For more information contact us by phone (using one of our toll-free numbers listed below) or by mail. Then cast your vote for...

SDC Search Service a division of
System Development Corporation
2500 Colorado Avenue, Santa Monica, CA 90406
Toll Free Numbers: (800) 382-6689 (CA)
(800) 421-7229 (continental U.S. except CA)

Manuscripts of articles should be addressed to William D. Mathews, Editor, *Journal of Library Automation,* NCLIS, 1717 K St., N.W., Washington, DC 20036. Technical communications and news items should be addressed to Mary A. Madden, *JOLA Technical Communications,* Blackwell/North America, 10300 S. W. Allen Blvd., Beaverton, OR 97005. Copies of books submitted for review should be addressed to Judith G. Schmidt, *JOLA Book Reviews,* 1408 D St., S.E., Washington, DC 20003. Advertising arrangements should be made with Susan K. Martin, *JOLA Advertising,* 134 Trinidad Dr., Tiburon, CA 94920.

Journal of Library Automation is the official publication of the Library and Information Technology Association, a division of the American Library Association, 50 E. Huron St., Chicago, IL 60611; *Executive Secretary:* Donald P. Hammer. The journal is issued quarterly in March, June, September, and December.

Journal of Library Automation is a perquisite of membership in the Library and Information Technology Association. Subscription price, $7.50, is included in membership dues. Nonmembers may subscribe for $15 per year. Single copies, $4.

Circulation and Production: American Library Association, 50 E. Huron St., Chicago, IL 60611. Please allow six weeks for change of address.

Publication of material in the *Journal of Library Automation* does not constitute official endorsement by the Library and Information Technology Association or the American Library Association.

Abstracted in *Computer & Information Systems, Computing Reviews, Information Science Abstracts, Library & Information Science Abstracts, Referativnyi Zhurnal, Nauchnaya i Tekhnicheskaya Informatsiya, Otdyelnyi Vypusk,* and *Science Abstracts Publications.* Indexed in *Current Contents, Current Index to Journals in Education, Education, Library Literature,* and *Quarterly Bibliography of Computers and Data Processing.*

Microfilm copies available to subscribers from University Microfilms, Ann Arbor, Michigan.

Second-class postage paid at Chicago, Illinois, and at additional mailing offices.

Cl**ing the Catalog

In the beginning, it was very simple. Some wit at one of the foundations had seen a Woody Allen film and dashed off a handwritten note to the Library of Congress. The note read, "Why don't you try closing the catalog, there might be big bucks in that." It wasn't long before wheels were rolling, the best minds were put on the problem, as they say, and the idea was run up the flagpole. A few of the people working on plans for a Shrine to the Dustjacket were reassigned and brought their talents to bear. "This should not just be a closing," one suggested, "this should be a *grand* closing: something to outdo all grand openings and all grand closings of supermarkets, shopping centers, and hotdog stands." "We'll have a ribbon cutting," proposed another, "perhaps ice cream and cake in the pavilion." "Get the Pentagon involved," cried a third, "have them airlift a national figure in by helicopter, we might get Melvil Dewey dressed like Santa Claus."

Bit by piece the plans fell into place. A year-long statistical study was commissioned with an outside consultant. This served to quantify the scope of the problem more precisely and seemed to show that, from a statistical point of view, closing the catalog was really much nearer at hand than many had suspected. On a drawer-by-drawer basis, the catalog was found to be entirely closed between 10 p.m. and 6 a.m. During the remaining sixteen hours of the day, there was a low-level clatter of drawers reaching a peak at 11 a.m. and another dull crescendo between 2 and 3 in the afternoon. But even at its height, 97 percent of the drawers were completely closed, while another 2 percent were only partially open and could be slammed closed at a moment's notice. Only a mere 1 percent were completely out of their slots, recklessly lying about.

But before long a debate ensued as to how to close the catalog and really keep it shut. "If we're gonna get big bucks for this, we better do it right." Here there was a wide variety of opinion. A few relied on authority and suggested, "Just put up a sign proclaiming that the catalog is closed per order of The Librarian. No one would dare touch it then." Most felt more decisive action was appropriate; many adhered to the notion of Elmer's glue. Still others wanted a frontal attack. "Why not just come in one night with a bunch of hammers and simply nail the silly thing shut." A variation on this was the phased approach: nail one section at a time beginning, for example, with the letter *U*. Finally, it was determined that in order to justify the expenditure of big bucks, a really

sophisticated methodology would have to be adopted. Cryogenic technology took the day. Quite by accident it had been found that if the catalog were chilled to three degrees above absolute zero, the drawers became hopelessly frozen in place. This technology had a further advantage in that it significantly arrested card decay. So the preservationists were very much pleased. At about this time, though, it was also discovered that the coils on the freezers were so large that a completely new building would have to be erected just to house the cooling machinery.

Construction progressed, and other tasks on the PERT chart were checked off as well. A spool of lavender ribbon was purchased at the GSA store, the Smithsonian sent over a gold-plated scissors from the Columbian Exposition, and a contract for imprinted balloons went out for competitive bid. A large wicker basket was procured to hold the expected big bucks, and the date was set. At first the freezing was supposed to happen during the Bicentennial, but the building wasn't finished in time. Probably just as well, there were a lot of trivial events to befuddle the mind that year. Then, in line with the Santa Claus theme, Christmas of 1979 was proposed. But everyone wanted to stay home that day, so it was put off till just after New Year. Wednesday, January 2, 1980, became the day to watch. Finally, the public was informed.

Onlookers grew nervous. Across the country, people couldn't believe that the LC catalog would really be frozen. They coined euphemisms to hide the awesome truth. "Melting," said some. "Thawing." "Opening," they nodded in unison. A few began to suspect the grim reality and named the date "Black Wednesday." By and by the foundation representative called LC and said, "I hear you're closing the catalog, what gives?" A lengthy conversation ensued, the people at LC eagerly explaining how nicely the concept had grown, the ribbons, the complexities of the freezing apparatus. Not least, a gentle reminder of the big bucks was delicately made. "Closing? I don't remember saying that," the representative replied. "Seems to me I wrote a note about *cloning:* lots of little catalogs all over the place. *That* should be worth some bucks. I got the idea from a movie I saw."

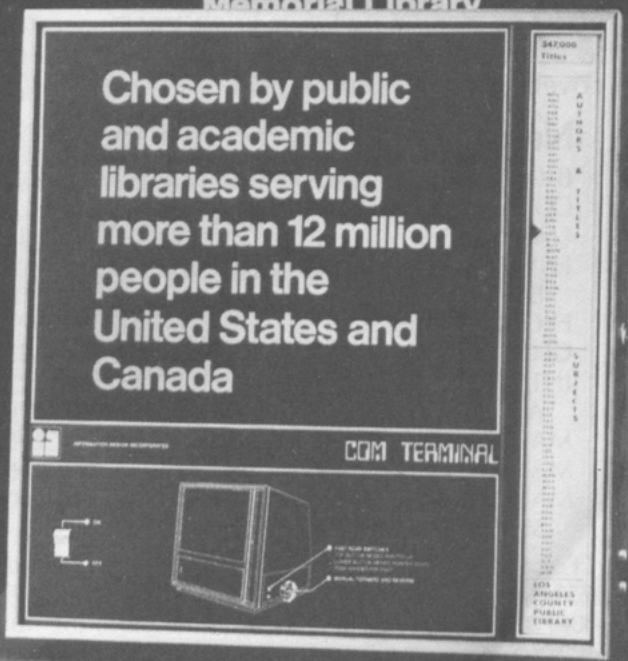
The world will end in fire or ice. The catalog? We can't be sure. But we *do* know that cloning isn't possible: the Congress would never stand for it in their very own Library right on Capitol Hill.

WILLIAM D. MATHEWS

The Standard of Excellence in COM Catalog Readers

Richmond Public Library
Cook County Sheriff's Department
Wichita Public Library
Raisin Valley Library System
Willard Library System
Bethel College
University of Toronto
Huntington Beach Public Library
Black Gold Cooperative Library System
San Luis Obispo City-County Library
Santa Barbara Public Library
Los Angeles County Public Library System
University of Southern California
Denver Public Library
Smithsonian Institution
Palm Beach County Public Library
Savannah Public Library
South Central Kansas Library System
Kansas State Library
Macomb Community College
Oakland Community College
Monroe County Public Library
J. J. Hill Reference Library
North Dakota State University
Duluth Public Library

St. Johns University
St. Paul Seminary
[illegible] Theological Seminary
St. Louis Community College District
Cumberland County Public Library
Lincoln University
Forsyth County Public Library System
North Dakota State University Library
Austin Public Library
Hidalgo County Library System
Houston Public Library
Arrowhead Library System
Mid-Continent Public Library
St. Louis County Library
Fairfax County Public Library
Prince William County Public Library
Marquette University Memorial Library



INFORMATION DESIGN
3247 Middlefield Road
Menlo Park Ca 94025
(415) 369-2962

University of Texas
North East Texas Library System

Our information retrieval system touches more bases than anyone.

The world's leading online retrieval service, Dialog now offers private and public libraries more than 75 data bases and more than 20,000,000 document references. No other retrieval service offers as many data bases or abstracts as Dialog. No other retrieval service covers as many subjects.

For libraries, Dialog also offers many other advantages. There is no minimum charge. You pay only for what you use. There are volume discounts for organizations making extensive use of Dialog. And there is the straightforward direct retrieval power of the Dialog language.

New Dialog bases recently loaded or forthcoming soon include:

BIOSIS PREVIEWS (1969-71 back files)
CONFERENCE PAPERS INDEX
ENERGYLINE
ENVIRONMENTAL PERIODICALS BIBLIOGRAPHY
EXCERPTA MEDICA
GEOARCHIVE
INTERNATIONAL PHARMACEUTICAL ABSTRACTS
MAGAZINE INDEX (popular magazines)
MANAGEMENT CONTENTS®
MARITIME RESEARCH INFORMATION SERVICE ABSTRACTS
MODERN LANGUAGE ASSOCIATION INTERNATIONAL BIBLIOGRAPHY
NIMIS (media for handicapped)
PAIS (Public Affairs Information Service)
PIRA (paper, printing, packaging)
SPIN (American Institute of Physics)
SSIE CURRENT RESEARCH
TOXIC SUBSTANCES CONTROL ACT CANDIDATE LIST
U.S. POLITICAL SCIENCE DOCUMENTS (Univ. of Pittsburgh)
WORLD TEXTILES

For a free catalog of data bases, write Lockheed Information Systems, Dept. 50-20, 3251 Hanover, Palo Alto, CA 94304. In the U.S., call toll-free (800) 227-1960. In California, call (800) 982-5838. For TELEX communication, the number is 346-409. For TWX, 910/330-9221.

Lockheed Dialog

A Word-Based Compression Technique for Text Files

Stephen F. WEISS and Russel L. VERNOR, III: University of North Carolina at Chapel Hill.

This paper presents a word-based technique for storing natural language text in compact form. The compressed text consists of a dictionary and a text that is a combination of actual running text and pointers to the dictionary. The word-based technique nearly halves the storage required with no loss in information. Furthermore, the process requires only a moderate amount of time overhead to compress and store the text and can retrieve and decode the encoded text faster than the original text can be retrieved.

INTRODUCTION

Computer-stored files of natural langauge text are becoming increasingly common. These files are used, for example, with programs for natural langauge analysis, information retrieval, and text editing. Such files are generally large, and users sometimes find themselves limited by the available computer storage space. If natural language text were optimally compact, we would simply have to live within the constraints. But language is highly redundant, and it would therefore be beneficial if some of this redundancy could be squeezed out and a more compact representation of the text stored. This paper discusses one such technique for compact storage of natural language text files.

The goal of this work is to develop a process for compression of text files that satisfies a number of constraints. First, we insist on perfect recall. The decoded version of the compressed text must be an exact duplicate of the original. Second, the process must be simple and easily programmed. And third, the compression and expansion processes must not involve excessive amounts of overhead in either time or space. Since a typical natural language text file is encoded and stored once and is then repeatedly accessed, we will be willing to tolerate a somewhat higher level of overhead in the compression and storage than in the expansion and retrieval.

Manuscript received July 1977; accepted January 1978.

Prior work in text compression falls roughly into two classes: *variable length encoding* and *inverted files*. Variable length encoding assigns a short code to the more common text elements and a longer code for the less common. Snyderman and Hunt[1] report a 35 percent reduction in the size of an upppercase text by encoding certain letter pairs in half the normal amount of space. Similar results have been achieved by Schieber and Thomas[2] and Maggs.[3]

The inverted file approach replaces the text with a string of pointers to a dictionary. Each pointer represents the segment of text to which it points in the dictionary. Compression results from the fact that the pointer is shorter than the text segment. A variation of the inverted file is the *partially inverted file*. Here, only some of the text segments are replaced by pointers, while the rest remain in unencoded form. Maggs[4] uses a partially inverted file based on letter pairs; Schuegraf and Heaps[5,6] and Thiel and Heaps[7] use equafrequent text fragments; Hahn[8] uses fixed-length text fragments; and Wagner[9] uses phrases to achieve compression. Heaps[10] uses an inverted file with variable length pointers; frequent words get short pointers, less frequent words get longer pointers.

Results of compression techniques are normally reported as the percent of the original text length that is saved due to compression. For the work cited above, the results are in the 25 to 45 percent range. In most of these studies, the texts contained only uppercase letters.

THE COMPRESSION PROCESS

The compression technique described here uses a partially inverted file and a dictionary of words. Long and/or frequently occurring words are placed in the dictionary, and each occurrence in the text is replaced by a pointer. Figure 1 shows a simple example of a word-based partially inverted file. The method is based on the IBM 360 architecture, which uses eight-bit bytes and the EBCDIC alphabet in which each character occupies one byte. The technique, however, is easily extended to other architectures and character codes. Our system can handle upper- and lowercase letters, numbers, and special symbols and has the capacity for 192 distinct characters.

We select words for the dictionary based on the amount of compression each word can produce. For example, if word W is L characters

Source text: It was the best of times; it was the worst of times.

Dictionary: 1 of
2 times
3 the
4 was

Compressed text: It 4 3 best 1 2; it 4 3 worst 1 2.

Fig. 1. Partially Inverted File Used for Compression.

long and occurs in the text with frequency F, and if pointers are each P bytes long, then the savings is

$$S = (L*F) - [(P*F) + D(W)].$$

The first term, $L*F$, is the amount of spaces occupied by the word in the unencoded text. The second term is the space used for the same word in the compressed text. It is made up of the amount of space needed for the pointers in the text ($P*F$) and the dictionary space required ($D(W)$). For short or infrequently used words, S might well be negative. Such words are not included in the dictionary, since their inclusion would lengthen rather than compress the text. Words for which S is positive are included in the dictionary and replaced in the text by pointers to the dictionary.

The Dictionary

For this work we used a dictionary with a maximum capacity of 1,024 words. We found that even large pieces of text containing several hundred thousand characters rarely had more than 1,000 words that offered significant savings. A longer dictionary would improve compression only minimally while increasing the time needed to encode the text.

We chose to use fixed-length dictionary entries. Fixed-length entries have the advantage of being far easier to access randomly than are variable-length entries. But they also have the disadvantage of wasting space for short words and not being long enough for long words. To minimize this effect, we use an eight-byte dictionary entry and allow a word to occupy either one or two entries. A bit associated with each entry signals whether it is a new word or the continuation of the previous word. This technique allows words of up to sixteen characters to be stored in the dictionary and, at the same time, keeps wasted space to a reasonable level. Approximately 25 percent of the dictionary words were longer than 8 characters, and hence the 1,024-entry dictionary could hold about 800 words.

The Pointers

In a partially inverted file, pointers must perform two tasks: they must identify themselves as pointers, and they must point to the dictionary entry for the word they represent. In the EBCDIC code, no printable character has a bit configuration beginning with 00. Therefore we use the leading double zero to identify pointers and distinguish them from characters. Our pointers are each two bytes long. Two of the 16 bits are for identification, 10 point to the dictionary (allowing a dictionary) of $2^{10} = 1{,}024$ entries), and the remaining 4 are used for other purposes explained below. One-byte pointers were tried, but with only 6 pointer bits ($2^6 = 64$), they could not handle a large enough dictionary.

We use one of the pointer bits as a capitalization flag. All words in the dictionary are stored with their first letters in lowercase. The capitalization flag tells whether or not the first letter of the encoded word is to be capitalized when decoded. This allows words such as "The" and "the," which differ only in the capitalization of their first letter, to be handled with a single dictionary entry and hence improves compression. No attempt is made to translate capital letters that occur in positions other than the first. If a letter inside a word is capitalized (as in "McDonald" or "USA"), it will be captialized in every instance of the word. Occurrences of "Mcdonald" or "Usa" would be extremely rare indeed.

One major disadvantage of the two-byte pointer is that it apparently will not allow savings to be realized for words of length one or two. This is not particularly serious for length one words; there are only two that occur regularly ("a" and "I"). But there are a great many common two-letter words, for example, "by," "to," "at," "or," "an," "on," "in," etc., and it would be highly desirable to be able to realize some compression from these. We can achieve this compression by noting that every word is followed by a delimiter, usually a space or a punctuation mark. By incorporating this delimiter into the pointer, we can realize one or more extra bytes of savings per word and thus achieve savings for words of length two or more.

The delimiter process is implemented using the three remaining pointer bits. Each of the eight bit configurations is associated with a possible word ending. Figure 2 shows the bit configurations and the associated endings. Since this set of endings is obviously incomplete, the empty ending (111) is provided, although it is needed only rarely. Thus for each word of length L, up to $L + 3$ characters may be represented by a single two-byte pointer. Figure 3 summarizes the pointer organization.

Creating the Dictionary

The first step in the compression process is creating the dictionary. This begins by calculating, for each word type in the text, the total savings that can be realized by its inclusion in the dictionary. For the spe-

Bit configuration	*Ending*
000	_ (space)
001	._ _ (period space space)
010	._ (period space)
011	,_ (comma space)
100	?_ _ (question space space)
101	;_ (semicolon space)
110	)_ (right paren space)
111	no ending

Fig. 2. Endings and Ending Codes.

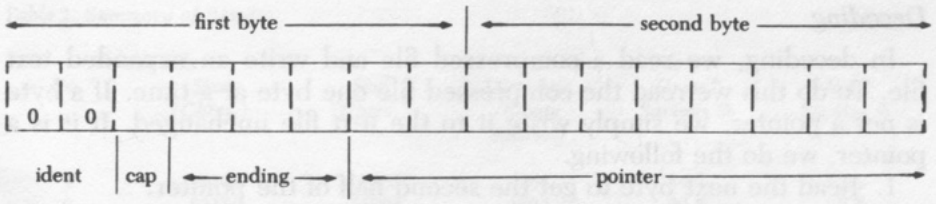


Fig. 3. Pointer Format.

cific values discussed in the previous two sections, the savings formula is:

$$S = (F*L) + E - [(2*F) + \lceil L/8 \rceil * 8]$$

E is the savings realized by compression of word endings. The $\lceil\ \rceil$ brackets indicate the ceiling function.

Next, the 1,024 words with the highest savings are sorted in decreasing order by savings. Then, starting at the top, as many of these words as will fit are placed into the 1,024-by-8-byte dictionary. The process essentially requires that the text words be sorted in order to bring together all instances of each word type. By using one of the efficient sorting algorithms, this process can be done in $0(n \log n)$ time where n is the number of word tokens in the text. The remainder of the process (picking the top 1,024, sorting them, and placing them into the dictionary) can be done with complexity $0(n)$ and thus the dictionary process is $0(n \log n) + 0(n) = (n \log n)$ for a text of n words.

Encoding

We will assume that we want to read an unencoded source file, compress it, and write a compressed file. We read the source file one word at a time and look up each word in the dictionary. If it is not found, the word and its trailing delimiters are written unchanged to the compressed file. If it is found, a pointer with the appropriate capitalization flag and ending code is written to the compressed file. The process requires $0(n)$ time, since each of n words must be subjected to some processing. The amount of time spent per word depends on the search algorithm. A binary search of 1,024 items can be done with only 11 comparisons. We made the search even faster by taking advantage of the fact that most words in the text are not found in the dictionary. We used a twenty-six-by-twenty-six array of bits; the (i,j)th element is 1 if one or more dictionary words has the ith letter of the alphabet as its first letter and the jth as its second. The (i,j)th element is 0 otherwise. Thus, for example, if element $(1,2) = 1$, we would know that at least one dictionary word begins with "ab." Conversely, if element $(1,2) = 0$, then the dictionary contains no words that begin "ab." By consulting the array before looking in the dictionary, many fruitless searches were avoided in only one step.

Decoding

In decoding, we read a compressed file and write an expanded text file. To do this we read the compressed file one byte at a time. If a byte is not a pointer, we simply write it to the text file unchanged. If it is a pointer, we do the following.

1. Read the next byte to get the second half of the pointer.
2. Retrieve the dictionary entry pointed to by the pointer bits.
3. Capitalize the first letter if the capitalization flag is set.
4. Write the word to the text file.
5. Write the indicated ending to the text file.

As with encoding, the time required by the process is a linear function of the text size. However there is no search involved; the dictionary is accessed directly, and the capitalization and ending require only a few steps each.

RESULTS AND ANALYSIS

Experiments with the compression process were performed on seven different text files. File one contained documentation for a software product; file two contained two chapters from a manuscript in natural language analysis. Both of these files were relatively short and were used primarily for debugging and initial testing. Files three and four each contain the text of ten articles in the area of population planning. The articles in file three are more or less specific; those in file four are general. Files five and six contain transcripts of lectures from a course in natural language analysis, and file seven contains two chapters from a manuscript in compiler design. An eighth experiment was performed using files five and six catenated together.

Table 1 summarizes the results of the eight experiments. For each file it shows the number of characters in the source and compressed file, the number of entries in the dictionary, the gross savings (dictionary not included), and the net savings (dictionary included). For all but the smallest file, the compressed file requires only about half the space of the original. When the space required for the dictionary is added in, performance degrades somewhat. For the small files this degradation is substantial, since the dictionary size is a sizable fraction of the file size. For the larger files, the degradation is only from about 3 to 7 percent. The variation in savings is due in part to file size. As the text size grows, so does the number of repeated words and hence the savings. However, since we are using a fixed-size dictionary, we may reach a point of diminishing returns when the text size becomes so big that a larger dictionary would be needed to take advantage of all the potential savings. We are probably seeing this begin to happen in the eighth experiment using the file of 324,000 characters. Some variation in savings is also due to stylistic differences from author to author, but among texts of approximately equal size, this variation is quite small.

Table 1. Summary of Results

Source File		Compressed File			
No.	*Size (Chtrs.)*	*Size (Chtrs.)*	*Dict. Size (8-byte Entries)*	*Gross* Savings %*	*Net† Savings %*
1	15100	9865	256	34.7	21.1
2	63150	31264	1024	50.5	37.5
3	147707	77879	1024	47.3	41.7
4	151791	82702	1024	45.5	40.1
5	122117	61015	1024	50.0	43.3
6	201893	100712	1024	50.1	46.1
7	138814	68368	1024	50.7	44.8
5+6	324010	167074	1024	48.4	45.9

*Gross savings excludes dictionary.
†Net savings includes dictionary.

Compression results alone are insufficient evidence to pronounce this system viable; it must also be computationally tractable. Therefore it is necessary to look at the time taken to perform the various steps. Quoting actual running times would be meaningless because it is highly machine and data dependent. Instead, we will compare the running time of each process on each file with the running time of a standard program using that same file. In our case, the standard program is one that reads a source file and writes it out unchanged. We will denote the time required for this program by t.

Creating the dictionary is the most expensive part of the process. It requires from $2t$ for the smaller collections to $2.5t$ for the largest. The coefficient of t will continue to grow slowly as the text size increases. The actual compression of the text requires approximately $1.3t$. The extra processing required to compress the text is partially offset by the decrease in the size of the output file and hence the relatively small overhead. The total dictionary creation and text compression process thus requires less than $4t$. This is not an unreasonably high value since the compression and storage of a piece of text need be done only once.

The overhead involved in retrieval of the compressed text is more important since we expect retrieval to be done repeatedly. Our results show that it requires approximately $0.85t$ to read the compressed file, decode it, and write the expanded file. Here the time saved due to the smaller input file more than compensates for the extra effort needed to decode. Thus we see that the compression overhead is not excessive and retrieving the compressed text is actually faster than retrieving the original text.

In addition to saving storage space and retrieval time, this process has two other advantages. First, the retrieval process is completely compatible with uncompressed files. If a file contains no pointers, the expansion algorithm will retrieve it unchanged. Thus only a single retrieval program would be necessary to handle a collection containing both

compressed and uncompressed files. Second, the dictionary provides a crude but useful indication of content. The dictionary contains mainly high frequency words. If function words (e.g., "the," "and," "to," "by," "is," etc.) are ignored, we have a list of high-frequency content words. And Salton[11] and others have shown that such words do very well in indicating the basic content of a document. For example, the top ten nonfunction words in the dictionary of the text on population planning were:

planning	clinic
family	patient
population	services
patients	contraceptive
health	program

Such content information is useful in content analysis, abstracting, and information retrieval, and it is available as a no-cost by-product of the compression process.

One task for which this process in its present form is not well suited is the compression of strings of blanks. Such strings do not generally occur in natural language text but are extremely common in programming language text. Indeed, fully half the symbols in a program source code may be blanks. Blank compression may be easily incorporated into our method in the following way. One pointer value, say, all zeros, would be reserved as a blank signal. Rather than point to the dictionary, it would indicate that a string of blanks had been compressed; the exact number of blanks would be contained in the byte immediately following the special pointer. Since this method would only be effective for compressing 4 or more blanks, a 0 in the following byte could indicate 4 blanks deleted, a 1 indicates 5 blanks deleted, . . . and 255 indicates 259 blanks deleted. Thus from 4 to 259 blanks could be compressed into only 3 bytes.

CONCLUSION

The word-based text compression has shown itself to be an effective text storage and retrieval technique. It nearly halves the storage required while still providing perfect retrieval of the stored text. It requires a tolerable amount of overhead for storage and actually saves time on retrieval. In addition it is compatible with uncompressed files, provides a useful, albeit crude, indication of text content, and is extendable to blank compression.

REFERENCES

1. M. Snyderman and B. Hunt, "The Myriad Virtues of Text Compaction," *Datamation* 16, no.16:36–40 (1970).
2. W. Schieber and G. Thomas, "An Algorithm for Compaction of Alphanumeric Data," *Journal of Library Automation* 4, no.4:198–206 (1971).

3. P. B. Maggs, "Compression of Legal Texts for More Economical Storage", *Jurimetrics Journal* 14, no.4:254–61 (1974).
4. *Ibid.*
5. E. J. Schuegraf and H. S. Heaps, "Selection of Equafrequent Word Fragments for Information Retrieval," *Information Storage and Retrieval* 9:697–711 (1973).
6. E. J. Schuegraf and H. S. Heaps, "A Comparison of Algorithms for Data Base Compression," *Information Storage and Retrieval* 10:309–19 (1974).
7. L. H. Thiel and H. S. Heaps, "Program Design for Retrospective Searches on Large Data Bases," *Information Storage and Retrieval* 8, no.1:1–20 (1972).
8. B. Hahn, "A New Technique for Storage and Compression of Data," *Communications of the ACM*, 17, no.8:434–36 (1974).
9. R. Wagner, "Common Phrases and Minimum-Space Text Storage," *Communications of the ACM*, 16, no.3:148–52 (1973).
10. H. S. Heaps, "Storage Analysis of a Compression Coding for Document Data Bases," *Infor* 10, no.1:47–61 (1972).
11. G. Salton, *Automatic Information Organization and Retrieval* (New York: McGraw-Hill, 1968).

Developing Corporate Author Search Keys

James W. BOURG: Gallaudet College; Douglas LACY: Cumberland County Public Library; James LLINAS: HRB-Singer; and Edward T. O'NEILL: State University of New York at Buffalo.

A systematic approach to the design of fixed-length, derived, truncated search keys for corporate author records is described. Certain distributional and informational characteristics of the elements of the entries are shown to be useful in search key design while some others are not. A statistical method is described for predicting the performance of search keys for files of arbitrary size. This method is employed for predicting the performance of a key known to perform well in a small sample file.

INTRODUCTION

The development of methods for the effective retrieval of bibliographic records from large data bases has received much attention from the library community with the advent of large on-line networks allowing resource sharing and thereby reductions in the costs of technical processing. The largest of these networks is the shared cataloging service operated by the Ohio College Library Center (OCLC), a system having approximately 1,400 member libraries, 1,600 network terminals, and a data base comprised of well over 3 million records. OCLC's approach to record retrieval has been to employ derived, truncated search keys for each of several index files through which the system provides access to the full bibliographic records. This approach has proved efficient for the retrieval of cataloging records by personal author, title, and personal author-title combinations.[1-3] However, retrieval problems have arisen during the application of this established technique of search key construction to records having corporate authors.[4,5]

In the case of corporate authors, one obvious approach to achieving retrieval improvement is through the use of a stop-list, through which some set of high-frequency words would be eliminated from considera-

Manuscript received June 1977; revised February 1978; accepted April 1978.

tion for key construction at the time of search key and index generation. Landgraf, Rastogi, and Long were the first to point out explicitly the existence of high-frequency words in this particular application.[6] They note that ". . . many corporate names begin with the same or similar words. For example, in the records examined, the initial words of more than 1,300 publications are 'U.S. Congress, House Committee on . . .'." This is a direct effect of the proper application of the *Anglo-American Cataloging Rules*. Rules 69 to 71 require entry of the record under the name of a hierarchically higher body, followed by the name of a subordinate body directly responsible for the publication. This situation occurs particularly often for government-produced publications. For these, rules 78 through 86 apply, but the general principle is that agencies through which the basic legislative, judicial, and executive functions of government are exercised should be entered as subheadings under the heading for the government (as in the Landgraf example above). Typical entries are therefore "posted upward" in hierarchy, with the result that the leading words of the entry are those associated with the higher, less specific, elements of the hierarchy. Since the keys from the corporate author words are constructed in a left-to-right manner, similarly high frequencies of occurrence of related search keys are realized, with an attendant degradation in record discrimination.

Recently, Smith and Rush[7] found that, for a given set of key-types, most keys were four times as specific for personal authors as corporate authors (in terms of key distinctness as used here—see later discussion). They report that a (4,1) key structure* operating on "U.S. Congress . . ." yields 10,000 replies for a file of about 68,000 records and also suggest the use of a stop-list of frequently occurring words. The OCLC system presently employs a thirty-one-word stop-list in the construction of corporate author search keys. These words are eliminated in a left-to-right fashion until the first significant word occurs at which point use of the stop-list ceases. Hence, subsequently occurring high-frequency words may enter into the key formulation. Although the actual system employs a stop-list in this fashion, no results have been published that compare the performance of keys with and without the use of a stop-list in the OCLC environment.

The purpose of the experiments reported here was two-fold: (1) to investigate various procedures for the construction of efficient truncated key-types as derived from corporate author records and (2) to evaluate the effectiveness of a representative key-type for a growing file. The methodology developed for the determination of search key-types and recommendations for specific key-types (for corporate author data) are seen to be applicable to operational systems such as OCLC, and practi-

*The notation indicates the number of characters employed in the key for each successive key candidate word in the record.

cal machine storage and human factor considerations necessary to such systems have been taken into account.

In the general case, construction of a truncated key-type for a set of records requires the development of three kinds of rules:

1. rules regarding which record elements are to be selected for key construction,
2. rules to determine the number of record elements to be selected for key construction, and
3. rules to determine which record element fragments are to be selected for key construction.

In particular, these rules may here be interpreted as defining (1) *which* corporate author words are to be used in search key construction, (2) *the number of words* to be used in corporate author search key construction, and (3) *the number of characters per word* to be used in the construction of the search key. At the outset, the only assumption regarding which words would be used in the keys (rule one) was that words would be used in a left-to-right fashion, i.e., truncated from the right. Except for this, no other prior assumptions on word ordering or word elimination were made.

The development of rules two and three determines the total key length in characters. Since the keys (and pointers) comprise the records of the index files,[8] considerations of efficiency of record storage often impact the choice of optimum key strategy. In the OCLC case, efforts are made to limit the total key length to eight characters so that one key requires at most two computer words of storage. Although the eight-character limit was not considered a strict boundary for key construction in this study, efforts were made to maintain the total number of key characters reasonably close to this value. While keys of up to twenty total characters were constructed, those that were considered as serious alternative keys for application in an operational setting (e.g., OCLC) were in all cases comprised of twelve or fewer total characters. Guidance for selecting the number of words and number of characters per word was sought primarily from certain distributional and informational characteristics of the record elements.

METHODOLOGY

Sample Data

The sample for the study was a file of 2,451 records consisting of corporate author entries extracted from the 110 field of a larger, 22,000-record sample of MARC II-format records of the OCLC data base. The larger test file was a systematically drawn sample from the entire OCLC data base at the time of this study (file size of approximately 10^6 records). Conference main entries (field 111) and corporate and conference names from other fields (410, 411, 710, 711, etc.) were not included in the sample.

Procedures

The approach to the study of corporate author search keys taken here involved the use of three different categories of measures. The first two categories of measures were used as an attempt to describe rules one, two, and three for key construction discussed above and to evaluate search key performance. A third category of measures was used to predict the performance of a given search key-type for a growing file.

Two measures in the first category were derived from the distributional characteristics of the sample corporate author entries; the mean number of words per record and the mean number of characters per word in the sample were used to establish rough upper limits for the number of words in a key-type and the number of characters to be used from the words entering into the key. Another measure used to determine the number of characters per word entering into the key was the conditional entropy of characters in corporate author words; this measure is used to quantify the informational value associated with additional characters in keys.

For the sake of a clearer intuitive understanding of this notion of entropy, consider the following example. Let us take a seven-character word and fill in characters one at a time, starting with the first character position:

L _ _ _ _ _ _
L I _ _ _ _ _
L I B _ _ _ _
L I B R _ _ _

The successive addition of characters has served to reduce the number of possible choices available for succeeding character positions (i.e., reduce the entropy), thus reducing the possible choices for word identity. If it is guessed that the intended word is LIBRARY, the remaining three characters (ARY) will have provided no new information (i.e., no additional entropy) and will be completely redundant.

A second category of measures was comprised of performance measures for the test key-types that were constructed, as computed for the sample file environment. Two measures were used, distinctness and relative entropy. Distinctness D is the ratio of the total number of distinct keys generated by a particular key construction strategy to the total number of distinct corporate author entries. Relative entropy H_r is the ratio of the entropy of a particular key-type to the total file entropy. Both of these measures quantify, in different ways, the average record discriminability associated with a given key construction strategy. The measures have been defined and used elsewhere in studies related to search keys.[9,10]

The third and final category of measures used were those employed

for predicting the performance of a given key construction strategy for the case of an increasing file size. To carry out these calculations essentially requires a procedure for estimating the characteristics of the record-to-key distributions for arbitrary file sizes. The method employed was originally developed by O'Neill[11] for the study of journal scattering and was modified for use in the current setting by O'Neill and Llinas.[12] The remainder of this section of the paper will be devoted to a more formal discussion of the methods employed for calculation of conditional character entropy, the problems associated with the calculation of word entropy, and the techniques developed for predicting key performance in growing files.

In the case of character entropy, calculations can be made that estimate the so-called "n-gram" entropy,[13] which is the conditional entropy of the n^{th} letter given the prior $(n\text{-}1)$-long character string. Shannon[14] has done this for written English. Bernard[15] has made similar calculations for French, German, and Spanish, and Newman and Waugh[16] have computed the redundancy (an associated measure) for Samoan and Russian. For very large n, the n-gram entropy F_n approaches the entropy H. Formally, we have:

$$F_n = -\sum_{i,j} p\,(b_i,j) \log_2 p_{b_i}(j) \qquad (1)$$

where: b_i is a string of $(n\text{-}1)$ letters
j is an arbitrary letter following b_i
$p(b_i,j)$ is the probability of the n-gram (b_i,j)
$p_{b_i}(j)$ is the conditional probability of the letter j following the string b_i

and $H = \lim_{n \to \infty} F_n$

Equation (1) produces a measure of the average uncertainty of the next letter j when the preceding $(n\text{-}1)$ letters are known. As n is increased, F_n accounts for statistics of longer strings and approaches the character entropy H. Since the strict calculation of F_n for English requires proper statistics of 26^n varieties of strings, approximate methods are often used to compute F_n for n larger than about 4. Values of F_1, F_2, and F_3 for free English text have been computed with relatively high precision[17] from tables of character to trigram frequencies.[18]

Newman and Gerstman[19] propose a method that estimates F_n by counting occurrences of character pairs separated by varying numbers of spaces in given character strings. The relation proposed is formally correct in certain limiting cases and, for the free English text sample used, yields values of F_n within about 10 percent of one of the Shannon[20] values for strings up to about ten characters in length. This procedure is used here to estimate the additional information provided by successive characters in words in corporate author entries.

Computation of word entropy as a means of determining the rule for the number of words for the key is, as has been discussed elsewhere,[21] an essentially impossible task. To do so in an accurate fashion would require a very large data base that would be sure to capture all "significant" words. Even using the approximate entropy-calculating method of Newman and Gerstman[23] requires that $m \times m$ contingency tables be determined for each of several interword spacings, where m is the total number of distinct words in the data sample; meaningful calculations therefore result in very large values of m. An estimate of the word entropy can be calculated by forming the product of the entropy per character and the mean word length in characters, although this assumes word-to-word independence, an unrealistic situation. Alternately, attempts have been made to use Zipf's law[23] for word probabilities to compute word entropy,[24] but it is well known that Zipf's law is not a proper probability distribution. This latter method also assumes word-to-word independence, whereas conditional word probabilities are required in the correct calculation. Thus, it is considered here that a rigorous approach is confounded by the existence of a limited sample, as well as by computer storage limitations for each of several very large arrays, and no calculations of word entropy are made.

The essence of the technique developed for predicting key performance in growing files is this: a statistical distribution is hypothesized to characterize the record-to-key rank-frequency distribution, estimates of the population values for the statistical parameters are obtained, and the resultant particular distribution is tested for goodness of fit with the sample data. This procedure having yielded satisfactory results, it is shown that a particular statistic, which depends on the population parameters of the hypothesized distribution and the file size, can be used to reexpress the performance parameters for any desired file size.

The particular distribution used is the negative binomal, which is a two-parameter distribution expressible as

$$p(n) = \frac{\Gamma(\rho + n - 1)}{\Gamma(n)\,\Gamma(\rho)}\,\phi^{\rho}\,(1-\phi)^{n-1} \qquad (2)$$

where $p(n) = \text{Prob}(x = n)$ for x, the true rank of a key, and ρ and ϕ are parameters. O'Neill[25] has shown that the traditional methods of parameter estimation yield biased estimates of ρ and ϕ but has developed a method by means of which unbiased estimates may be computed. The mathematical development for this method involves a statistic that depends on ρ, ϕ, and the file size. The performance parameters D and H_r can be reexpressed in terms of this statistic, so that with reliable estimates of ρ and ϕ, the performance can be computed as a function of file size. In addition, it can be shown that the cumulative probability of retrieving a set of records of a given size or less can also be expressed in

terms of this statistic, and this parameter is also computed as a function of file size. (This probability parameter has been used explicitly in earlier OCLC studies of key effectiveness.[26]) It has been shown[27] that the method produces reliable key performance predictions from small sample data for files nearly two orders of magnitude larger than the sample file.

RESULTS AND DISCUSSION

Selecting Candidate Words from the Record (Rules 1 and 2)

The distribution of words per record for the sample of corporate author entries is skewed and has a mean of about six words per record and a standard deviation of about three words per record. Ninety percent of the records have about ten words or less. Thus, allowing for ten candidate words would provide good file coverage and probably good discrimination, but it is felt to represent too much of a compromise in index file storage requirements (e.g., half the sampled records have five or fewer words). As a starting point, using the mean value would seem to be a reasonable choice for an upper word limit in terms of storage demands, file coverage, and probable specificity of record coverage (as yet unknown).

To pursue this issue of determining the number of words and characters in a key, an exploratory parametric study of key performance was carried out for keys using from two to four words, taking from one to three characters from each word. The particular key-types in this control group experiment were:

2 words: (2,2), (3,3)
3 words: (1,1,1), (2,2,2), (3,3,3)
4 words: (1,1,1,1), (2,2,2,2), (3,3,3,3)

The key performance results are shown both in terms of distinctness and relative entropy in figures 1(a) and 1(b) as the control group data. Both graphs indicate that the gains in performance obtained in adding third or fourth words to the keys are substantial and require further study of keys with more than four words. For any given number of words, the gains in performance in adding the second character per word are notable, with lesser improvements in adding third characters. However, even the best key performance is very low for such a small sample file.

The effect of a stop-list was considered next. A word-frequency count was performed on the entire sample file for this study in order to determine the candidate words for the stop-list. Most words occurred less than ten times. The obviously troublesome words were those occurring more than eighty times or in at least 3.25 percent of the records. A list of these words (list A) is shown in table 1. Also shown is a shorter list (list B), chosen for comparison with the first, made up of the 10 most popular words (which occur more than 160 times or in more than 6 per-

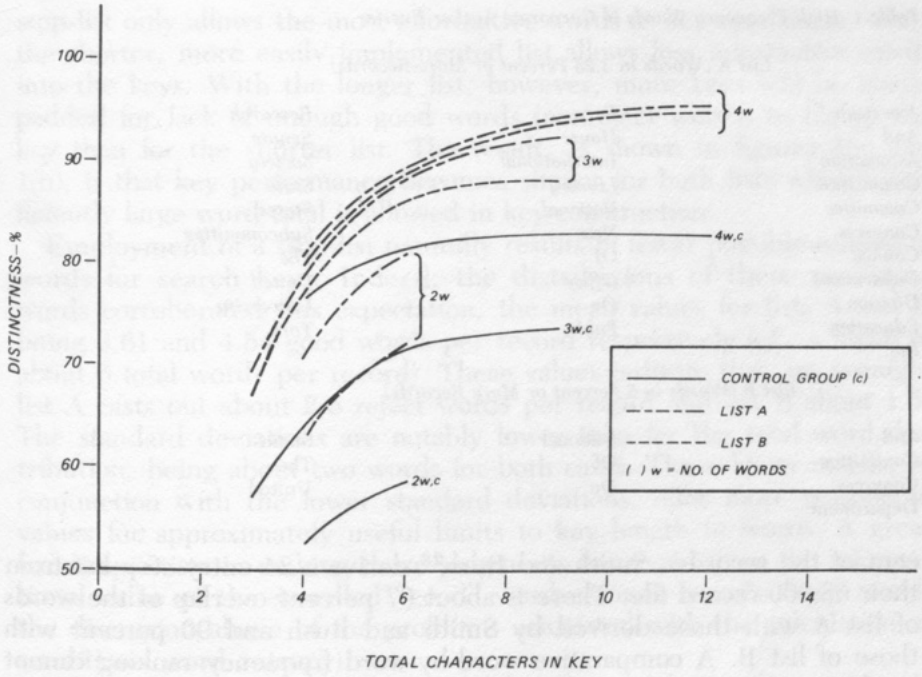


Fig. 1(a). Search Key Performance–Distinctness.

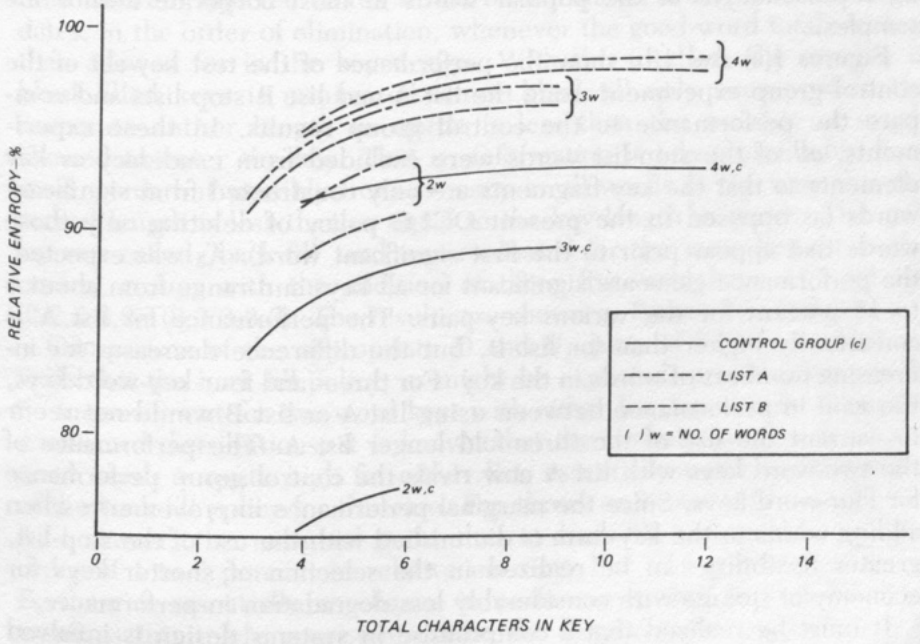


Fig. 1(b). Search Key Performance–Relative Entropy.

Table 1. High Frequency Words in Corporate Author Entries

List A (Words in 3.25 Percent or More Records)

American	*For*	*Research*
And	*House*	*Senate*
Association	*International*	*Society*
Commission	*Library*	*State*
Committee	*National*	*States*
Congress	*New*	*Subcommittee*
Council	*Of*	*The*
Department	*Office*	*United*
Division	*On*	*University*
Education	*Public*	*York*
Etc.		

List B (Words in 6 Percent or More Records)

And	National	States
Committee	Of	The
Congress	On	United
Department		

cent of the records). Smith and Rush[28] derive a 24-entry stop-list from their 68,000-record file. There is about 67 percent overlap of the words of list A with those derived by Smith and Rush and 90 percent with those of list B. A comparative word-by-word frequency ranking cannot be done, since specific word frequencies or rankings were not available in the Smith and Rush study. However, lists A and B certainly should be representative of the popular words in most corporate author file samples.

Figures 1(a) and 1(b) show the performance of the test key-set of the control group experiment using the list A and list B stop-lists and compare the performance to the control group results. In these experiments, *all* of the stop-list words were excluded from candidacy as key elements so that the key fragments are only constructed from significant words (as opposed to the present OCLC policy of deleting only those words that appear prior to the first significant word). As was expected, the performance gains are significant for all keys and range from about 5 to 15 percent for the various key-pairs. The performance for list A is consistently higher than for list B, but the difference decreases for increasing numbers of words in the key. For three- and four-key-word keys, the gain in performance between using list A or list B would not seem to warrant the use of the three-fold longer list A. The performance of the two-word keys with list A now rivals the control group performance for four-word keys. Since the marginal performance improvements when adding words to the key-form is diminished with the use of the stop-list, greater flexibility can be realized in the selection of shorter keys for economy of storage with considerably less degradation in performance.

It must be realized that a compromise in systems design is involved with the selection of particular stop-lists. On the one hand, the longer

stop-list only allows the most informative words as key candidates, while the shorter, more easily implemented list allows less informative words into the keys. With the longer list, however, more keys will be blank-padded for lack of enough good words (nonreject words) to fill up the key than for the shorter list. The result, as shown in figures 1(a) and 1(b), is that key performance becomes similar for both lists when a sufficiently large word total is allowed in key construction.

Employment of a stop-list naturally results in fewer possible candidate words for search keys. Indeed, the distributions of these candidate words corroborated this expectation, the mean values for lists A and B being 3.61 and 4.54 good words per record respectively (*cf.*, a mean of about 6 total words per record). These values indicate that, on average, list A casts out about 2.5 reject words per record and list B about 1.5. The standard deviations are notably lower than for the total word distribution, being about two words for both cases. These lower means, in conjunction with the lower standard deviations, offer more reasonable values for approximately useful limits to key length in words. A great handicap exists in making these judgments however, since the informational value of the so-called "good" words cannot be calculated (recall the discussion above of the problems associated with the calculation of conditional word entropy).

Taking up the policy that some information is better than none (as provided by blank key-fragments), an alternative strategy was considered wherein rejected words would be reinstated as key element candidates, in the order of elimination, whenever the good-word total was less than allowed for in the key design. With this method, the number of blank-filled keys is minimized, since blank-fill only occurs when the corporate author length in words is less than the number of words allowed to enter the key. That is, information is taken preferably from good words but possibly from as many as all words in the case that the record length is less than or equal to the key length in words. This policy was called "back-fill" and was evaluated for the same key-set as the control group, with the addition of three five-word keys: (1,1,1,1,1,), (2,2,2,2,2), (3,3,3,3,3). The results are shown in figures 2(a) and 2(b).

Comparison of the appropriate D and H_r plots of figures 1 and 2 reveals that the back-fill policy, as might be expected, provides essentially no improvement for two-word keys with either list A or list B. From the good-word distributions, it is known that the probability of at least two good words is quite high for either list, essentially nullifying the effect of the back-fill policy for this case. For three- or four-word keys with list A, however, the marginal gains due to back-fill are notable, on the order of 1 to 3 percent in relative entropy. For all key-forms using list B, there is essentially no change in performance due to back-fill. This result is intuitively consistent, since the keys constructed with list B admit more of the less-informative words (although labeled good words)

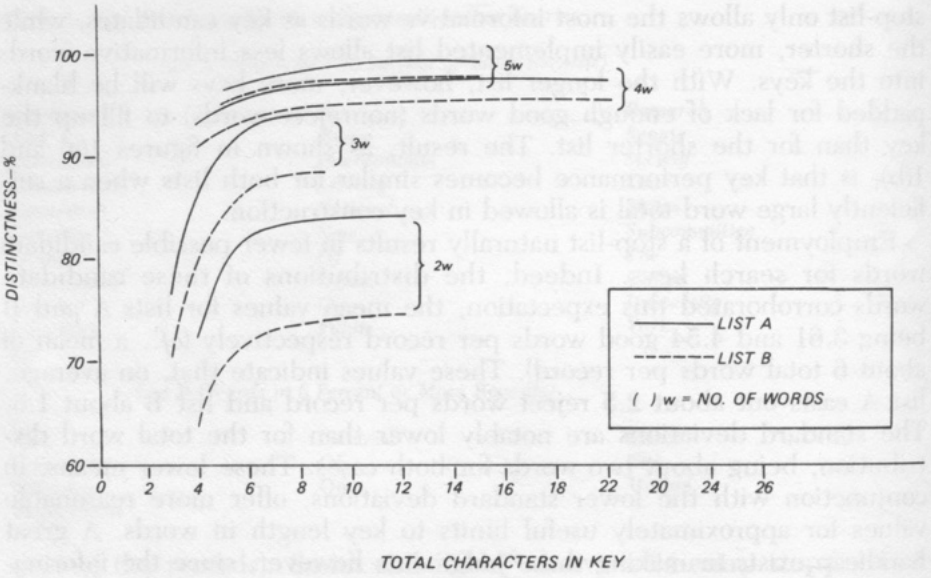


Fig. 2(a). Search Key Performance with Back-Fill–Distinctness.

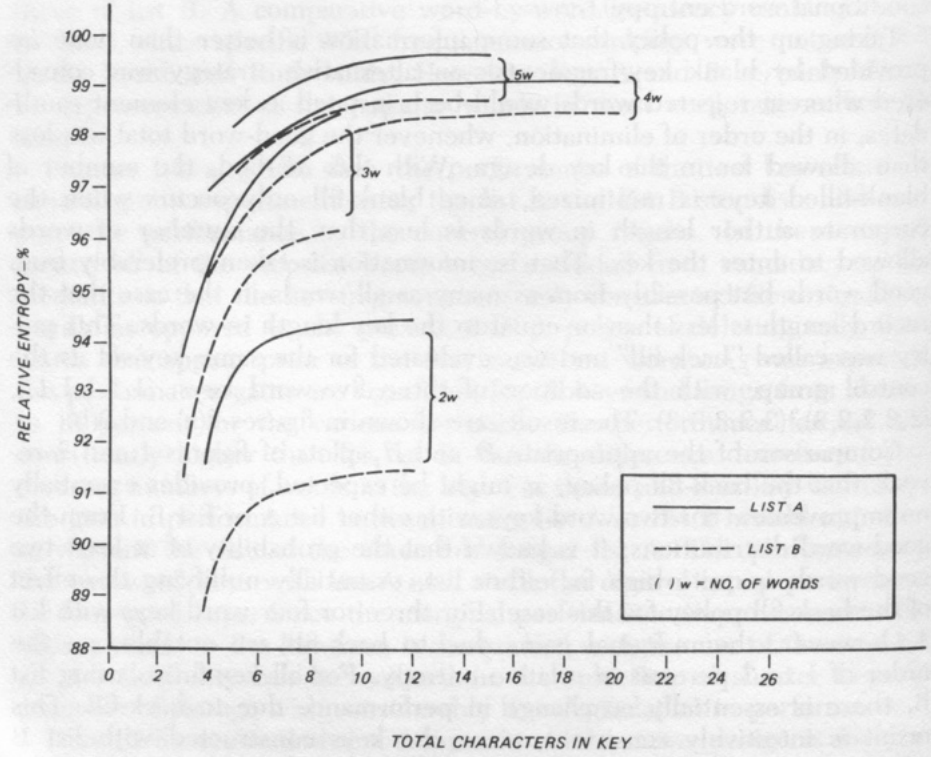


Fig. 2(b). Search Key Performance with Back-Fill–Relative Entropy.

than do the keys for list A. The marginal performance gains of the added five-word keys compared to the four-word key performance are on the order of ½ to 1 percent, and so the utility of adding words beyond the fifth seems limited. On the basis of these results, five-word keys were taken as the upper word limit for key construction.

Figure 2 also shows the results for the additional key types (5,5) and (6,6), which were added to clarify the asymptotic trends of the two-word key performance curves. An added (5,5,5,5) key shows that essentially no change in performance occurs with the use of such high total character keys.

In summary, we may make the following conclusions regarding the word selection procedure for these corporate author search keys:

1. A stop-list of some type should definitely be employed.
2. The useful range of words in keys is from three to five words, and in this range the performance differences between the use of long and short stop-lists is relatively small.
3. The use of the back-fill policy results in slight performance increases and only with long rejection lists; for a short rejection list, no change in performance should be expected, even for as many as five words in the key.
4. The mean values of the good-word distributions for either stop-list correlate reasonably well with the optimum number of key words associated with the parametric performance calculations and may be used as measures to limit the useful length of keys in words.

Selecting the Number of Characters/Key Element (Rule 3)

The mean number of characters per good word for each of lists A and B were determined as about seven characters per word, with standard deviations also on the order of seven characters for both cases. Ninety percent of the words are about ten to eleven characters or shorter in length.

The mean number of characters per good word, however, cannot be used as a rough estimate for the upper limit on characters per word-in-key; while yielding good performance, it represents an excessively lengthy key strategy. The performance data suggest a value in the neighborhood of three to four characters per word-in-key, which corresponds to the points at which the slopes of the performance curves become particularly flat.

Calculations of the conditional character entropies in good words in corporate author entries were made here with the technique described above. The results are shown in figure 3. Since the lists A and B differ only by twenty-one words, the computations yield essentially the same values of F_n for both lists, since the very large character sequence is little changed by the absence of those words. As expected, the values of F_n display an asymptotic behavior, trending toward H for increasing

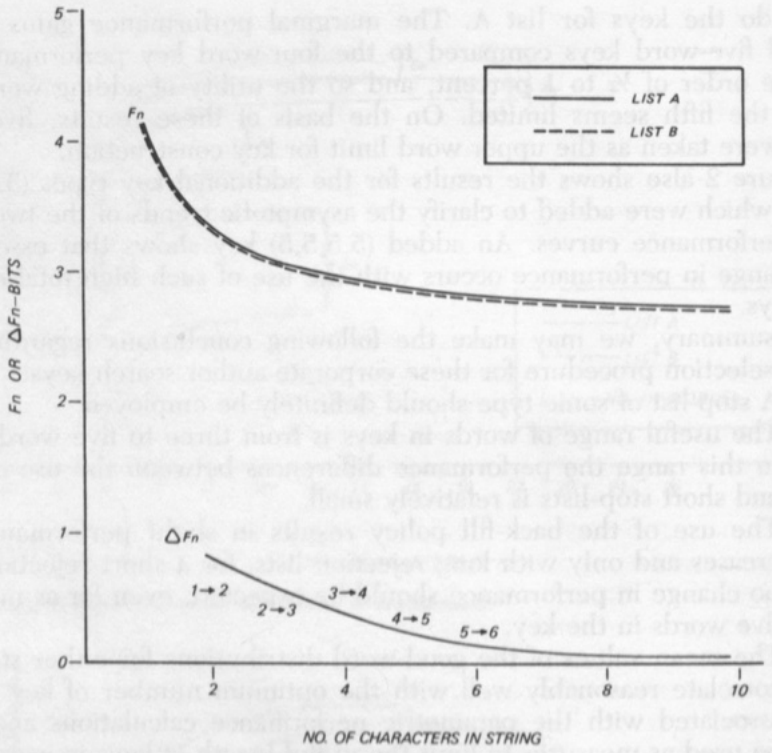


Fig. 3. Conditional Character Entropy Estimates.

character string length. The limiting value of H here is approximately 2.7 bits per character. For free English text, Shannon[29] computed values of 2.62 and about 2 bits per character using different methods, whereas Newman and Gerstman[30] computed a value of about 2 bits per character for a different sample.

Shown also on figure 3 are the changes in F_n for increasing character string length. These changes are quite small beyond a key element length of about four characters and indicate that the information added by characters beyond the fourth in a given string is quite small. It is clear from this graph that: (1) the greatest proportion of character information resides in the first few characters and that (2) the useful character limit (three to four characters) suggested by these data agrees reasonably well with the useful limit observed in the performance curves of figures 1 and 2.

The above discussion is not intended to suggest that the problem of the number of words to be included in a key and the number of characters to be included in each word (key element) are independent problems. This is not the case. When looked at in their entirety, the key performance experiments demonstrate that, for a given total number of

key characters, keys with more words and therefore fewer characters per key element perform better than keys with fewer words and more characters per key element. Thus, there is a complex interaction between the number of words (key elements) and the number of characters per key element. Moreover, this study has been limited to a uniform number of characters taken from each candidate word. It is of course possible that the information value of good words is also position dependent and that varying the number of characters per word for each position will alter the overall ranking of the various strategies.

Performance Predictions for Varying File Sizes

The estimation procedure for ρ and ϕ (parameters of the negative binomial distribution discussed above) involves a rather complex computer-based trial-and-error procedure. Therefore, a parametric approach to the file-dependent performance calculations for the search keys is infeasible and in fact not necessary. Calculations need only be carried out for a key whose performance in the sample file has shown promise. The key chosen from the set evaluated in figure 2 is the (2,2,2,2,2) key using list B and back-fill. This key was felt to represent the best choice if an actual implementation were to be considered. However, since back-fill results in minimal improvements (*cf.* figures 1 and 2), this policy need not be used in the face of implementation problems (e.g., the human factors consideration of the difficulty of key formulation for data base searching in an operational system such as OCLC).

The rank-frequency distribution of records to keys for the (2,2,2,2,2) strategy is shown in table 2. For the sample file used here, the most popular key was associated with 21 records, whereas 1,746 keys were associated with only one record. In all, 1,995 distinct keys were produced.

These data were used to estimate the values of ρ and ϕ to determine the population characteristics of this key-type and compute the expected

Table 2. Rank Frequency Characteristics for (2,2,2,2,2) Corporate Author Key (List B, with Back-Fill)

Number of Keys	*Number of Records*
1	1746
2	164
3	44
4	17
5	11
6	6
7	2
10	1
13	1
14	2
21	1
Total Records: 2451	Total: 1995

performance parameters for varying file sizes. The values produced by the estimation procedure as described above were 0.56801 for ρ and 0.0001543 for ϕ. Chi-squared was computed as 13.72 and, for 5 degrees of freedom, is not significant at 0.01 significance level. Figure 4 shows the cumulative probability values for the observed data (table 2), the negative binomial or population values, and the expected sample distribution. There is very good agreement between the expected sample and actual sample data.

Figure 5 shows the entropy, relative entropy, number of distinct keys, and distinctness as a function of file size for file sizes ranging to about 10^6 records. All parameters show a strong dependence on file size. However, key entropy displays an asymptotic behavior for file sizes above about 10^5 and achieves a maximum value on the order of 13.1 bits. Distinctness drops drastically for larger files, since this particular key strategy produces distinct keys at a much slower rate than that at which records are added. Relative entropy changes gradually at first but then, as the entropy becomes asymptotic, displays, as expected, a slope corresponding to the slope of an inverse logarithmic relation.

Shown also in figure 5 are the probabilities of getting thirty-two or fewer, eight or fewer, and one or fewer records in response to an arbitrary key submitted to the system. For file sizes on the order of 10^6 records, all these probabilities are below 20 percent. Thus, in practical terms, with this key design, it would be rather likely that many users would need to employ an extended search procedure and page through several screenfulls of citations for files exceeding 10^6 records, as is the case with OCLC.

These calculations show that it is the average information content of the given key-set, i.e., its entropy, that achieves a stable value independent of file size for the larger files. This measure may therefore be used as a global measure of efficiency for a particular key-type and be used to compare the efficiency of various key-types. For the system designer, however, the more practical measures of distinctness and probability of a given number of records per key are probably of more direct utility. These parameters depend on file size in a nonlinear fashion but may be estimated from the procedures used herein.

CONCLUSIONS

An effort has been made to approach the problem of corporate author search key design in a systematic fashion using fixed-length keys for all records and a small sample of corporate author data. Computational difficulties prevent the calculation of the conditional information associated with successive words in a string, and so the determination of the useful length in words of a key construction technique must be done in an empirical, parametric fashion. However, when a stop-list is used, distributional analyses of the number of good words in records also yield

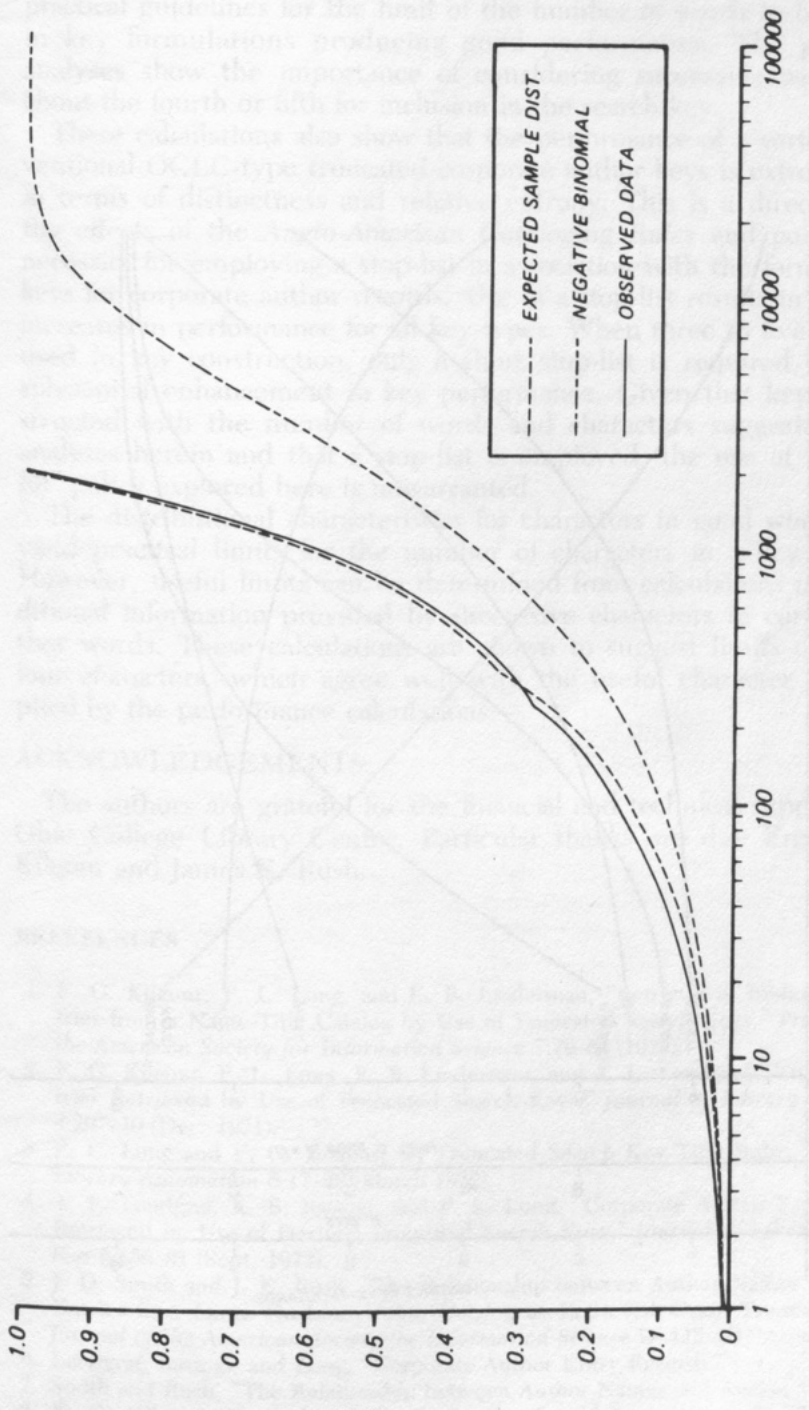


Fig. 4. Cumulative Probability Distributions ((2,2,2,2,2) Key, List B, with Back-Fill).

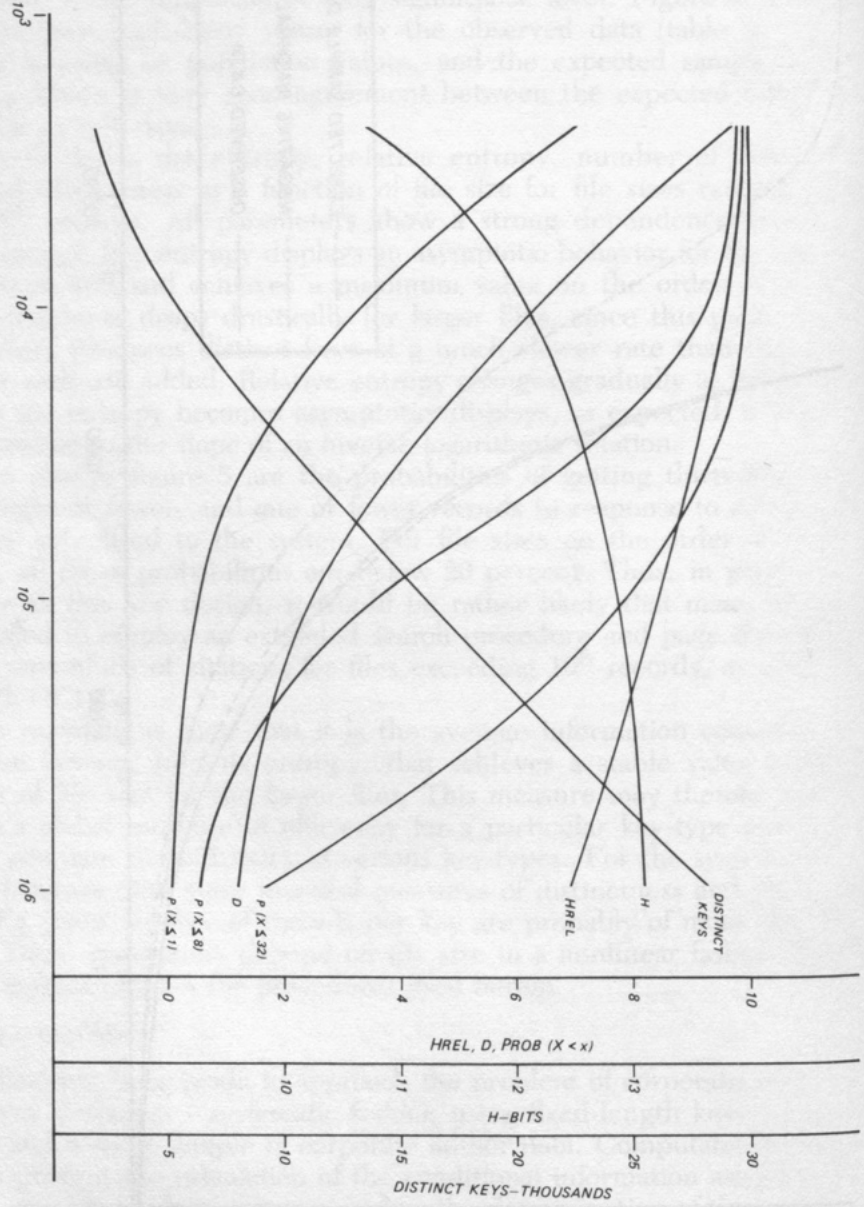


Fig. 5. Predicted Performance Measures for Arbitrary File Sizes ((2,2,2,2,2) Key, List B, with Back-Fill).

practical guidelines for the limit of the number of words to be included in key formulations producing good performance. The parametric analyses show the importance of considering successive words up to about the fourth or fifth for inclusion in the search key.

These calculations also show that the performance of a variety of conventional OCLC-type truncated corporate author keys is extremely poor in terms of distinctness and relative entropy. This is a direct result of the effects of the *Anglo-American Cataloging Rules* and points to the necessity for employing a stop-list in association with the formulation of keys for corporate author records. Use of a stop-list results in significant increases in performance for all key-types. When three to five words are used in key construction, only a short stop-list is required to achieve substantial enhancement in key performance. Given that keys are constructed with the number of words and characters suggested by the analyses herein and that a stop-list is employed, the use of the "back-fill" policy explored here is unwarranted.

The distributional characteristics for characters in good words do not yield practical limits for the number of characters in a key fragment. However, useful limits can be determined from calculations of the conditional information provided by successive characters in corporate author words. These calculations are shown to suggest limits of three to four characters, which agree well with the useful character limits implied by the performance calculations.

ACKNOWLEDGEMENTS

The authors are grateful for the financial and technical support of the Ohio College Library Center. Particular thanks are due Frederick G. Kilgour and James E. Rush.

REFERENCES

1. F. G. Kilgour, P. L. Long, and E. B. Liederman, "Retrieval of Bibliographic Entries from a Name-Title Catalog by Use of Truncated Search Keys," *Proceedings of the American Society for Information Science* 7:79–82 (1970).
2. F. G. Kilgour, P. L. Long, E. B. Liederman, and A. L. Landgraf, "Title-Only Entries Retrieved by Use of Truncated Search Keys," *Journal of Library Automation* 4:207–10 (Dec. 1971).
3. P. L. Long and F. G. Kilgour, "A Truncated Search Key Title Index," *Journal of Library Automation* 5:17–20 (March 1972).
4. A. L. Landgraf, K. B. Rastogi, and P. L. Long, "Corporate Author Entry Records Retrieved by Use of Derived Truncated Search Keys," *Journal of Library Automation* 6:156–61 (Sept. 1973).
5. J. D. Smith and J. E. Rush, "The Relationship between Author Names and Author Entries in a Large On-Line Union Catalog as Retrieved Using Truncated Keys," *Journal of the American Society for Information Science* 28:115–20 (March 1977).
6. Landgraf, Rastogi, and Long, "Corporate Author Entry Records."
7. Smith and Rush, "The Relationship between Author Names and Author Entries."
8. F. G. Kilgour, P. L. Long, A. L. Landgraf, and J. A. Wyckoff, "The Shared

Cataloging System of the Ohio College Library Center," *Journal of Library Automation* 5:157–83 (Sept. 1972).
9. A. L. Landgraf and F. G. Kilgour, "Catalog Records Retrieved by Personal Author Using Derived Search Keys," *Journal of Library Automation* 6:103–8 (June 1973).
10. D. W. Fokker and M. F. Lynch, "Application of the Variety-Generator Approach to Searches of Personal Names in Bibliographic Data Bases—Part 1. Microstructure of Personal Author's Names," *Journal of Library Automation* 7:105–18 (June 1974).
11. E. T. O'Neill, "A Stochastic Scattering Model," *Proceedings of the American Society for Information Science* 11:155–9 (1974).
12. E. T. O'Neill and J. Llinas, "A Method for Evaluating Search-Key Performance," *Proceedings of the American Society for Information Science* 13:42 (1976).
13. C. E. Shannon, "Prediction and Entropy of Printed English," *Bell System Technical Journal* 30:50–69 (1951).
14. *Ibid.*
15. G. A. Bernard, "Statistical Calculation of Word Entropies for Four Western Languages," *IRE Transactions on Information Theory* IT-1:49–53 (March 1955).
16. E. B. Newman and N. C. Waugh, "The Redundancy of Texts in Three Languages," *Information and Control* 3:141–53 (1960).
17. O'Neill, "A Stochastic Scattering Model."
18. F. Pratt, *Secret and Urgent, the Story of Codes and Ciphers* (New York: Blue Ribbon, 1942).
19. E. B. Newman and L. J. Gerstman, "A New Method for Analyzing Printed English," *Journal of Experimental Psychology* 44:114–25 (1952).
20. Shannon, "Prediction and Entropy of Printed English."
21. Newman and Gerstman, "A New Method for Analyzing Printed English."
22. *Ibid.*
23. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (New York: Addison-Wesley, 1949).
24. Shannon, "Prediction and Entropy of Printed English."
25. O'Neill, "A Stochastic Scattering Model."
26. Landgraf and Kilgour, "Catalog Records Retrieved by Personal Author."
27. J. Llinas, "A Method for Evaluating Search-Key Performance" (unpublished doctoral dissertation, State University of New York at Buffalo, 1977).
28. Smith and Rush, "The Relationship between Author Names and Author Entries."
29. Shannon, "Prediction and Entropy of Printed English."
30. Newman and Gerstman, "A New Method for Analyzing Printed English."

A Study of Errors Found in Non-MARC Cataloging in a Machine-Assisted System

Cynthia C. RYANS: Catalog Department, Kent State University Library, Kent, Ohio.

One of the most important impacts in cataloging in recent years is the use of a cooperative cataloging data base. In order to make the most advantageous use of such a system, it must contain uniform and accurate cataloging. This article reports the results of a study of cataloging records input into the OCLC data base by participating libraries (non-MARC records). The records were analyzed to determine the various fields of the cataloging record where most errors occur as well as the types of errors that are being made on the cataloging input into the data base.

INTRODUCTION

Libraries throughout the country are now using cooperative cataloging and have been doing so for many years. The Library of Congress *National Union Catalog* is a prime example of this. The most recent breakthrough in cooperative cataloging is OCLC, a computerized cataloging data base designed to increase the availability of academic library resources.

In order for library cataloging to be consistent in any cooperative cataloging system, basic standards of cataloging should be followed. Many libraries throughout the country are now using the *Anglo-American Cataloging Rules* (*AACR*) and the various chapter revisions of these standards, which include the International Standard Bibliographic Description for cataloging material received by their libraries.[1,2]

Described in this article are the results of a study designed to determine what areas of descriptive cataloging vary the most from the

Manuscript received June 1977; revised February 1978; accepted April 1978.

above-mentioned standards. The data in this study are based on cataloging that has been entered into the OCLC data base.

Since there are more than 1,200 libraries now participating in the OCLC system, it was felt that this system would provide an appropriate base from which to collect a variety of cataloging examples from many different sized libraries and a variety of special library cataloging.[3] This study in no way reflects on the quality or standards of OCLC. It should be realized that there is always some cataloging that does not meet the specified standards. The results of this study point out the most frequent errors. This knowledge should reduce the necessity for making corrections on catalog records and/or identify for users the fields where errors are most likely to occur.

OCLC STANDARDS

The OCLC Advisory Committee on Cataloging held a series of meetings during 1971 and 1972 to draft basic cataloging rules for all participating OCLC libraries to use when inputting cataloging into the OCLC system. In order to maintain high quality of input cataloging, a document was drafted by the Advisory Committee on Cataloging. This document was directed at making the records on OCLC compatible, so that these records could be used by all libraries with a minimum of change while at the same time conforming to Library of Congress cataloging and *AACR*.

This document, entitled "Standards for Input Cataloging," states ". . . that LC practice should be followed in so far as the practice can be determined by a local library following its own policies as to verification."[4] The document further states that the *AACR* should be used when the Library of Congress practice cannot be determined. The OCLC document lists those fields or subfields that are mandatory, required if available, optional, and those that should be omitted.

Since *AACR Chapter 6 (Separately Published Monographs)*[5] incorporates *ISBD(M)*[6] rules for monographs and the Library of Congress is using these revised rules, the OCLC Advisory Committee on Cataloging issued a document recommending that effective 1 September 1975 it be mandatory for OCLC participating libraries to apply these revised rules for cataloging monographs.[7]

With this background on the basic OCLC rules for input cataloging, a study was conducted of the OCLC participating library copy (non-MARC) to determine what fields were being omitted from cataloging and where variations from the basic rules were being made.

METHODOLOGY

We felt constrained to design an experiment that could be integrated into the everyday workload of the cataloging department. Therefore, the

records analyzed for this study comprised the first 700 monographic records input into OCLC after 1 September 1975 to which Kent State University Library attached its holdings as part of the normal processing activity. While it may be argued that the sample could be biased and might not be characteristic, we know of no partciular reason to expect that such bias would be significant. We feel, then, that the findings are strongly indicative of the data base as a whole.

Since the purpose of this study is directed at analyzing the catalog records by fields, no attempt was made to categorize by any special type of cataloging (open entries, title entries, corporate authors, etc.). The records studied all had been entered into the data base by an OCLC participating library as original cataloging. No Library of Congress records are included in the study. Serials and nonbook materials (i.e., films, phonodiscs, etc.) were not included.

Since OCLC required that all libraries use the revised *AACR* as well as the new ISBD punctuation effective 1 September 1975, only those records entered into OCLC after this date were considered in this study.

Each record was analyzed by field. A field, as defined by OCLC, is "a collection of data elements treated as a single unit. Each field possesses a name describing the contents of the field, e.g., imprint."[8] The fields that were checked in this study include main entry, title statement, edition statement, imprint, collation, series, subjects, and added entries. Neither notes nor call numbers were taken into consideration in this study. There are a number of reasons for varying call numbers—in some instances, several different call numbers can each be correct for the same title; some libraries use a class-together number for a series, while others class the series separately; Cutter numbers must vary in order to fit into each library shelflist, etc.

Notes were omitted because in many cases they vary according to the book in hand, for example, missing pages, autographed copy, etc. Many notes are the standard bibliography or index notes: "prepared for" notes, "first published" notes, "reprint" notes, etc., which rarely generate errors.

CORRECT RECORDS

For the purpose of this study, the concept of a "correct" record was developed. A "correct" record will be referred to as one that follows the *AACR* and *ISBD(M)* rules and the OCLC "Standards for Input Cataloging." This is the standard that the OCLC Advisory Committee on Cataloging requires in order to make the cataloging in the data base uniform. An "incorrect" record is considered to be any piece of cataloging that deviates from this standard. No attempt was made to determine local variations in cataloging procedures of participating libraries or Kent State University Library.

The results of the study indicate that basically the cataloging is of a very high quality. Of the 700 records studied, 417 (or 60 percent) were correct and could be accepted with no changes. This is a very important aspect of cooperative cataloging, since its purpose is to have cataloging available that can be used by everyone who wants to use it with no changes necessary. It should be pointed out here that many of the records analyzed for this study had only very minor errors, but nevertheless they were counted as errors, thus making the total number of correct records lower. However, in many cases, they could have been considered as "correct" records. An example of this would be in the added entry field where the inputting library omitted a joint author.

The total number of errors found in this study was 393. These 393 errors were found on 283 catalog records checked for this study. In some cases, several errors were found on one piece of cataloging. However, in 160 instances, there was only one error on the catalog record. Thus, of the 283 records with errors, 56 percent required only a minimum of change.

MAJORITY OF ERRORS

The two fields on the catalog record that generated the majority of errors were the subject heading and the collation fields. In the subject heading field, errors were found in eighty-eight records, or 31 percent of those records found to have errors (table 1). These errors take a vari-

Table 1. Errors Listed by Field

Fields	*Number of Errors*	*Percent*
Main Entry	44	16
Title	12	4
Remainder of title	3	1
Responsibility	13	5
Edition Statement	4	1
Imprint		
Place of publication	10	4
Publisher	5	2
Date	1	.4
Added Publisher	2	.6
Printer	1	.4
Collation		
Pagination	20	7
Illustration	21	7
Size	70	25
Series	55	19
Subject Headings	88	31
Added entries	44	16

700 records studied
417 (60 percent) correct
393 errors found on 283 records

ety of forms. For example, there were some catalog records where no subject headings were assigned at all. This can be a library preference, as some libraries do not assign subject headings. However, according to the OCLC "Standards for Input Cataloging," the subject heading field is required.[9] Therefore, those records on which subject headings were omitted were considered incomplete.

On seventy-five of the eighty-eight records with errors in the subject heading field, the subject headings were either incorrect or incomplete. On the remaining thirteen records, the subject headings were omitted. One source of error was the fact that the subjects were not Library of Congress subject headings. Other examples were found in which subdivisions in subjects were used incorrectly, such as the incorrect use of "indirect" or "direct" in subject headings where a place was involved.

Collation

The second field in terms of number of errors was the collation field. In the collation field, seventy records, or 25 percent of those records found to have errors, omitted the size of the book. As shown in this study, many libraries feel that this is not necessary to add to the catalog record. However, the *ISBD(M)* standards state that the size is an element of the collation field, and the collation field is a required field in a catalog record.[10] The *AACR* also states, "The collation is the cataloger's description of the physical work and consists of a statement of the extent of the work in pages, volumes, or volumes and pages, the important illustrative matter, the size, and accompanying materials, if any."[11] According to the "Standards for Input Cataloging" prepared by OCLC, entering the size on a record is required if available.[12] Therefore, omission of the size on a catalog record was considered an error.

Main Entry

The main entry field and the added entry field also generated quite a few errors. These two fields are somewhat related in this study, because if a book is not entered under an editor but rather entered under the title, then the editor must be traced. Therefore, when a change is made in the main entry field, it also affects the added entry field. This was found in a number of the records studied. In the main entry field, forty-four errors were found or 16 percent of those records containing some type of error (the entry was incomplete, the piece was entered under the incorrect entry, etc.). In the added entry field also, forty-four errors were recorded. Similar discrepancies were found in both fields, such as misspellings in the names, incomplete names, and dates either incorrect or omitted. Some of these errors simply fall into the category of typing errors. But in the case of corporate entries and corporate added entries, several records were found where these entries were in-

complete. As with other fields in the catalog record, there were errors in these two fields in both spelling and capitalization.

Remainder of the Catalog Record

The fields mentioned above contained the majority of the errors found in this study. However, the remaining fields in the catalog record did contain various kinds of errors worth mentioning. For example, in the title field, twelve errors (4 percent) were found. The major problem here was capitalization. The *AACR* states that "in the titles of books, pamphlets, periodicals, documents and other publications, legal cases, and works of art, the first word is capitalized. The capitalization of the other words in the title is governed by the other rules for capitalization."[13] Other records showed words were omitted, and a number of records revealed spelling errors.

In the remainder of the title subfield, only three errors (1 percent) were found. In these cases, the remainder of title was either omitted or incomplete. The area of responsibility subfield showed thirteen mistakes, or 5 percent of those records with errors. In the majority of cases the person or body responsible for the work was omitted or the information was incomplete.

In only four records (1 percent) was there any indication of an error in the edition statement. In three of the records the edition was incorrect, and in the fourth record the edition statement was omitted.

In the imprint field of the catalog record, the place of publication subfield contained the most errors. Ten errors were found in this subfield, or 4 percent of those records with errors. In the majority of these cases, as well as in the five records where errors were found in the publisher statement, the place or publisher was either omitted or incomplete. Only one record showed a discrepancy in the date of publication.

In the collation field, as stated earlier in this study, there was a high percentage of errors. Twenty errors (7 percent) were found in the paging, and 21 errors (7 percent) were found in the recording of illustrations. To further explain the errors in these two subfields, the paging errors were mostly incorrect recording of paging, possibly some typing errors, whereas the majority of errors in the illustration statement was the omission of the fact that the book contains any illustrations. A pattern appeared here, where in many cases the paging was given but nothing else in the collation statement. However, as mentioned earlier, the paging, illustrative matter, and size are all part of the collation and must be included in the collation statement. Therefore, in this study, this field was recorded as incorrect if the illustrative matter and size were omitted.

The final field in the catalog record to be discussed is the series statement. In fifty-five cases, or 19 percent of those records that contained errors, the series statement was incorrect or incomplete. The

errors ranged from the omission of the series and the omission of the series number to spelling errors. In some cases the subdivision in corporate entries in the series statement was omitted.

SUMMARY

Basically the cooperative cataloging analyzed for this study was relatively free of errors. This is very important in any cooperative cataloging system. It saves time for the cataloger as well as making the records more uniform.

The field that contained the most errors was the subject heading field. This is an extremely important field in a catalog record, and its correctness is critical when assigning a call number. The field with the second highest number of errors was the collation field. Although every field on a catalog record is important, this perhaps is the one field where the cataloging is still usable without completing the field.

Many of the errors found in the 283 records containing them were spelling or typing errors. These are minor errors but nevertheless make the catalog record incorrect.

It should be mentioned here that there are several possible reasons why the above-mentioned cataloging errors are being made. Perhaps the main reason is the lack of knowledge of the rules. The *AACR* and *ISBD(M)* standards are relatively new and do take time to both understand and incorporate into the cataloging of new titles. Another possible reason for the errors mentioned in this study could be administrative decisions on certain cataloging policies. As mentioned earlier, the catalog records in this study were not examined by library, but perhaps administrative policy on cataloging could be determined if such a study was made. Another reason some of the errors were made in the catalog records studied was simply carelessness. This can be verified by the errors found in spelling and perhaps some of the errors in the paging.

Overall, the quality of the cataloging checked was very good, and in the 700 records analyzed in this study, the errors that were found were relatively minor. This indicates that libraries are following the *AACR*, *ISBD(M)*, and OCLC "Standards for Input Cataloging" standards very closely and are trying to make a cooperative cataloging system both accurate and useful to other institutions.

CONCLUSIONS

In order to have a good, workable cooperative cataloging system, all the cataloging must be uniform. OCLC is attempting to make the system uniform by requiring a set of cataloging standards for all participating libraries. Since these standards have been made mandatory, the uniformity of the cataloging has improved. However, as shown in this study, there are still some libraries that are not cataloging according to the prescribed standards set up by OCLC. If all institutions in any

cooperative cataloging system would follow a prescribed set of standards, the system would be even more valuable and all records could be accepted as readily as Library of Congress copy is accepted.

REFERENCES

1. *Anglo-American Cataloging Rules, North American Text* (Chicago, American Library Association, 1967).
2. International Federation of Library Associations, *ISBD(M), International Standard Bibliographic Description for Monographic Publications* (London, IFLA Committee on Cataloging, 1974).
3. "OCLC Participating Libraries: Arranged by OCLC Symbol" (unpublished document, Ohio College Library Center, July 1977).
4. "Standards for Input Cataloging" (unpublished document, Ohio College Library Center, 2 June 1972), p.1.
5. *Anglo-American Cataloging Rules, North American Text: Chapter 6, Separately Published Monographs* (Chicago, American Library Association, 1974).
6. International Federation of Library Associations, *ISBD(M).*
7. "AACR Revised Chapter 6 and ISBD(M)," (*On-line Cataloging,* Addendum No.1 [unpublished document, Ohio College Library Center, July 1973]), p.1.
8. Ohio College Library Center, *On-Line Cataloging* (Columbus, Ohio: Ohio State University Libraries, Office of Educational Services, 1973), p.3.
9. "Standards for Input Cataloging," p.12–15.
10. International Federation of Library Associations, *ISBD(M),: Chapter 6, Separately Published Monographs,* p.22.
11. *Anglo-American Cataloging Rules, North American Text,* p.47–48.
12. "Standards for Input Cataloging," p.9.
13. *Anglo-American Cataloging Rules, North American Text,* p.348.

An Inside Look at the Copyright Office Cataloging System (COPICS)

David E. PULLMANN: Library of Congress, Copyright Office, Washington, D.C.

An automated system has been developed for cataloging the monographs, periodicals, music, maps, films, recordings, and realia received by the Copyright Office for registration. Some of the problems in disseminating this unique data base through national networks are considered. The system features certain special techniques—automatic inversion of personal names and the use of one entry as a model for the next.

INTRODUCTION

When sound recordings became registrable for copyright protection on February 15, 1972, an automated system was born that was to change life considerably for the Cataloging Division of the Copyright Office. Since the creation of a new registration class provided the opportunity to build a complete data base from the inception of the data itself, it seemed reasonable to capture the catalog data in machine-readable form from the beginning. What began then as a pilot project for a single class was soon extended to almost all classes of materials registered. The system, which became known as COPICS (Copyright Office Publication and Interactive Cataloging System), was designed by Leo J. Cooney, then systems analyst of the Copyright Office and later chief of the Cataloging Division.

BIBLIOGRAPHIC OR LEGAL RETRIEVAL?

From the beginning, a basic design question emerged, which was to persist throughout the development of the system. Should COPICS be designed merely to store and retrieve enough data to search a copyright claim or should it consider the broader interests of scholarly research as well? The scholar, no doubt, would wildly wave the pennant for access to this unique data base yet, once on-line to it, might sift through a mountain of dross to discover few nuggets of any value whatsoever.

Manuscript received February 1977; revised February 1978; accepted April 1978.

Obviously, there is a point of diminishing returns. Of the half million items registered annually, nearly 200,000 are unpublished. While this statistic is not necessarily pejorative, it frequently includes amateur and vanity works: Aunt Tillie writes a song for the next Dahlia Society pageant and registers her copyright "jest in case." To the student of folk art, however, a vein of musical works of this ilk is a treasure of great value. (A survey of song titles alone would yield some interesting results; consider, for example, such certain hits as "I'll Be Lovin' You 'til the Statue of Liberty Squats," "Who Put the Benzedrine in Mrs. Murphy's Ovaltine," "Whale Breath," and "Downwind of a Camel.") Moreover, the early works of an artist are occasionally learned of only through the unpublished registrations of his youth (e.g., those of humorist Ring Lardner and a number of composers).

On the other hand, *publication* is by no means a suitable criterion for determining which kernels are likely to satisfy scholarly appetites, either. Of 100,000 "books" registered each year, about 40,000 are hardcover monographs. Much of the remainder are rules for games, assembly instructions for garage doors and barbecue grills, self-styled mimeographic materials, and other ephemera of doubtful interest even to the Vertical File Index. But within this class also are University Microfilms registrations, a great deal of educational materials, and a wide variety of brochures, pamphlets, and propaganda for which Copyright Office records may represent sole bibliographic control. Whether published or not, many items registered are intrinsically of dubious advantage to the scholarly community, aside from purely commercial interest: lamp base designs, fabric prints, homemade jewelry, piggy banks, plaques, greeting cards, birdbaths, picture frames, and commodity packaging labels.

Primarily, of course, the Copyright Office must be concerned with the protection of the exclusive rights to property registered by its remitters, but, as an agent of Congress, it is also mandated by the Constitution "to promote the progress of science and the useful arts . . . ." The key issue is whether the bibliographic flow of copyright entries into the networks of the nation is promoting progress or simply sludge in the system.

COPYRIGHT CATALOGING

It is inconceivable to a librarian that material received for cataloging should *not* be intended for retrieval. Yet this is almost precisely the case for the Copyright Cataloging Division. Material is received as property to record a claim to ownership; the property itself, once recorded, is not only secondary in importance, in a sense, but is not necessarily available through the Copyright Office, the Library of Congress, or even the owner.

Although the tracing of the claim to ownership is of supreme impor-

tance in all considerations of retrieval, the item itself must still be described in sufficient detail to properly identify it. Why not, say some, while the item is in hand, catalog it "properly"—according to recognized bibliographic standards—and achieve the best of both worlds? A worthy objective, few would deny, but the great diversity of material deposited, both domestic and foreign, presents some problems: standard cataloging rules can be readily applied to most literary, law, cartographic, serial, audiovisual, educational, and even sound recording matter, but what should we do about the five yards of calico identified by the manufacturer only as Pattern No. 8914, a T-shirt reading "Kiss me, I'm Irish," an obscene greeting card, or the large economy-size bottle of Brand X bath oil? This seems to require some liberties with *AACR*.

Choice of entries and establishment of authority files are especially delicate matters, because each registration involves the legal names of authors and claimants. Sources of transcription are not only the title page, preliminaries, colophon, and the like, wherever conventional rules apply, but also include the information supplied on the application form for copyright registration as well. The cataloging of in-analyticals is governed less by rules for anthologies than by the remitter's option to register either the individual works or the collection as a whole. And so on. While acceptable compromises can be made for many of these kinds of problems, the mission statement of the Copyright Office formulated by Congress to establish copyright rather than bibliographic records is controlling.

THE MANUAL SYSTEM

Under manual methods, catalog entries were typed on photo-offset mats and processed through printing equipment to produce a number of three-by-five cards for each entry. From the cards, three separate catalogs were developed:

- a massive copyright card catalog (CCC), normally current to within a few weeks, containing entries dating from 1898 and, in some instances, earlier;
- a semiannual book *Catalog of Copyright Entries (CCE)* in nine parts—books, periodicals, dramas and speeches, music, maps, visual arts, commercial prints and labels, motion pictures and filmstrips, and sound recordings—produced by shingling the cards and photoduplicating the large boards (pages) upon which they were mounted by hand;
- a service that issues cards directly to subscribers who deal with high-volume copyright matters and place great emphasis on currency.

The "main entry" of each record is the registration number, which is always prefixed by a classification letter derived from the registration categories enumerated by §5 of Title 17, U.S.C. (e.g., *A* for books, *E*

for music, *N* for sound recordings). Title, all authors (including editors, arrangers, and sometimes illustrators), and claimants are the only other access points. In the *CCE* these are indexed in the back of the volume and refer to the appropriate registration number under which the main body of the catalog entry is given. For many years no series titles were traced, but this policy was discontinued in 1974. Unfortunately, no subject approach is possible.

The primary vehicle for dissemination of the cataloging data has been the *CCE*. Copies are available from GPO and are also distributed through the deposit libraries system. As increasing registrations began to outstrip its publishing schedule, however, the Copyright Office found it difficult to be timely in issuing its nine semiannual volumes. When lag times reached two years and longer, the effectiveness of the book catalog as a tool for legal research was little short of useless. Automation had been discussed as early as 1969, but initial efforts failed to produce a viable model. Now the time seemed ripe to address this issue once again.

THE ON-LINE SYSTEM

The automated system was designed to issue the products the Copyright Office had always issued in the past. That is, the *CCE* is still published, the Copyright Card Catalog is still maintained, and subscribers continue to receive cards of cataloged entries. (Unlike the MARC system, cards are not photocomposed; rather, special card stock is run through an ordinary impact printer in a weekly batch process.) In addition, the existence of the machine-readable record opens a variety of new formats: a microfiche or microfilm version of the *CCE* using COM equipment is in final testing stages, some segments of the *CCE* are now marketed directly in magnetic tape, and just over the horizon is the development of a retrieval system via CRT, which will include networking to subscribers throughout the world.

The individual entries of the *CCE* are created in an on-line, real-time environment. The task is accomplished with some eighty CRT terminals attached by dedicated phone circuits to the computer at the Library of Congress four miles away. To ensure accuracy, COPICS is divided into three primary functions: CATALOG (the creation of the entry itself), REVISE (a mandatory review process by a person other than the cataloger), and REVIEW (an optional review step at the supervisory level).

The Catalog Function

Each piece received for cataloging has been stamped with a unique registration number, and the cataloger begins his function by transmitting a transaction code containing the registration number. The system responds by returning to the CRT screen a list of four-letter mnemonic

abbreviations, called "prompts" (similar to MARC "tags"), each corresponding to a specific element of the entry—e.g., CLNA for claimant name, TITL for title, PHYS for physical description, etc. Fields are not subclassified nor subdivided, so that the indicators and subfield codes we are accustomed to seeing in MARC records are not needed.

Included in the list of prompts are a MESG prompt, reserved strictly for intraoffice remarks concerning the entry, and a XREF prompt for making cross-references, which generates a separate card when the entry is printed in card format. Space is also provided on the screen for reentering the registration number to assure that the entry data corresponds to the proper number; a discrepancy between this number and the one used in the initial transaction code results in a rejection by the system when a transmission of the completed entry is attempted.

Technical Solutions

To keep keystrokes to a minimum, added entries may be enclosed, *as they lie* in the entry descriptive paragraph, in braces: {}. For added entries that do not lie in the entry in the form in which they are to be traced, a special ADEN prompt is available, which allows virtually any number of added entries to be listed. In the case of personal names, a special delimiter within the braces may be used (presently a grave accent mark) to trigger an automatic inversion by the computer, e.g.: {John Q. Doe}. The use (or misuse) of the grave accent mark is deserving of an aside. It will be less puzzling and perhaps seem less uncouth in light of the primary objective of the cataloging function to create a data base for the purpose of determining ownership of a copyright. Therefore, any diacritics that do not affect filing are dropped; those that do are simply spelled out (i.e., oe for ö, aa for å, etc.). Thus, the search for a keyboard with a hatch mark or some other neutral symbol to use as a delimiter is unnecessary.

Many filing problems involving compounding, spacing vs hyphenation, and the like were eliminated without elaborate programming by the decision to simply adopt character-by-character sorting of filing terms. Certain other difficulties, such as roman numerals and abbreviations, were resolved through the cataloging rules or by directive. In character-by-character sorting, however, identical surnames may become separated, "Greene, Alfred" falling between "Green, Antonio" and "Green, Edward." To keep like surnames together, a special delimiter is used. The cataloger simply doubles the comma (or left parenthesis in the case of a corporate name beginning with a personal name) between surnames and forename. For example, {Doe,, John Q., Jr.} or {Doe ((John Q.) Marketing Company}. Of course, for the automatic inversion described earlier, no further treatment is necessary beyond placement of the grave accent to ensure proper filing. If any portion of the term within braces is to be excluded altogether from filing, it may be en-

closed within paired angle brackets. This is used mostly for foreign articles, e.g.: {<Die> Fledermaus}.

COPICS Edit

When the cataloger transmits the completed entry to the computer, the data is "edited" and rejected back to the CRT screen when certain conditions are present:

1. A misspelled date in the DATE prompt, e.g., 10Jal76.
2. Unpaired braces.
3. An improper added entry, e.g., if the collation (PHYS field) were put in braces.
4. A cross-reference in the wrong format, e.g., if the word SEE is omitted.
5. An angle bracket found not within braces.

After creating data fields for all appropriate prompts, the cataloger has several options:

1. CONTINUE, which allows him to reach prompts not found on the first screen.
2. START, which permits him to "page" back to the first screen after a CONTINUE.
3. NEXT, which stores the entry and automatically gives him the list of prompts for the next sequential registration number.
4. HOLD, which places the entry in temporary storage for later recall.
5. END, which stores the entry and allows a new transaction.

Model

Frequently, where material is received in sets, individual entries may differ from each other only in certain particulars, such as pagination or volume number. Here the cataloger may elect to construct a single entry and recall it over and over again as a MODEL for each of the subsequent entries, saving countless keystrokes in the process. Often registration numbers are in sequence, and great strides can be made by combining the MODEL feature with the NEXT option.

The Revise Function

After the entry is completed with an END or NEXT option, the cataloger may recall the entry at any time to make adjustments and modifications but only until the REVISE function begins. In REVISE, the entry is viewed from a broader vantage point—that is, the reviser scrutinizes the entry not only for discovering mechanical mistakes but, more importantly, for its comprehensive representation of a registered work. In deference to this principle, a reviser has been defined as one who knows when to break the cataloging rules.

How much "revising" is actually done is a matter of discretion. The reviser follows one of two courses: First, the reviser has free access to

the data on the screen and may correct any problem that lies there. Secondly, the reviser may REJECT the entry back to the cataloger with a note in the MESG prompt explaining the problem. When an entry is rejected, a code is affixed to the prompt so that only the data for that prompt may be changed by the cataloger. This, of course, obviates re-revision of the entire entry when it is transmitted back to the reviser.

The Review Function

When the reviser completes the entry with an END or NEXT option, it becomes available for the final step in the on-line processing—REVIEW by the supervisor. REVIEW stands in exactly the same relationship to the REVISE function as REVISE to CATALOG except that it is optional. If an entry is not called up in REVIEW by the end of the week following the week in which it is *revised*, the entry is automatically removed from the on-line file and stored on magnetic tape. A reviewed entry is removed and stored off-line at the end of the same week it is reviewed. One enterprising supervisor discovered that by using the NEXT option he could set an aspirin bottle on the XMIT key of his terminal and review entries flashing before him at a comfortable pace without ever touching the console.

Editing and Publishing

The tape that accumulates the records as they are processed off-line at the end of each week is the master file used in *CCE* production. The records to be compiled in each of the nine catalogs are first extracted on separate tape files. A hardcopy proof is then generated from each extract tape and sent to a text editor. Errors found in the proof copy are corrected via CRT in the following manner: The editor simply enters a correction transaction code using the registration number of the entry to be corrected, and a dummy record for that entry appears on the screen. After determining which prompt is associated with the incorrect data field, the editor then rekeys a new field with the corrected data for this prompt in the dummy record. When all corrections have been made in this manner, the dummy records are matched to the extract tape, and the new corrected fields in each record are substituted for the corresponding incorrect fields existing on the extract tape.

The updated extract tape may then be printed as camera-ready copy and sent to the Government Printing Office for binding and publication. Finally, the extract data is merged with a cumulative *archive* file, the permanent and primary data base of COPICS. If necessary, this archive file may also be updated through an on-line transaction similar to that for the extract tape.

PERIODICALS

The intrinsic nature of a periodical as a single unit whose parts cannot be assembled at one time precluded the recording of works registered

in class B, *Periodicals*, in COPICS at the time of implementation. Instead, a separate module for handling them was developed and implemented in February 1976.

In the manual system, periodical titles were maintained on four-by-six cards in a series of long, open card file trays. As an issue was registered, a cataloger-recorder pulled the appropriate title card and "posted" the issue data in pencil on the appropriate line beneath the title. Before long, cataloger-recorders were stepping over each other in reaching for cards. So it was necessary to sort the periodicals by title, and each cataloger-recorder was assigned responsibility for a portion of the alphabet and so a portion of the card file. At *CCE* time a complete set of new title cards was typed up for use in recording the new catalog period, while the entire series of cards for the ending catalog period was sent to a typing staff for preparation of the large boards (pages) from which the *CCE* was produced.

In the automated system each cataloger-recorder establishes the title of a periodical on a magnetic disc the first time it is received by entering a transaction code for this purpose at the CRT terminal. Thereafter, each issue received is recorded by retrieving the record established under its title and entering the appropriate data for that issue. When an attempt is made to establish a record for a title that is already "on file," the system responds with a query asking whether a duplicate title is intended (for occasionally two or more periodicals have identical titles). If the answer is "yes," the new record is established; if "no," the record for the existing title is displayed on the screen.

In retrieving a record to add new issues, time, frustration, and keystrokes are saved by using a short form known as a 3111 code. Already familiar to users of OCLC and some other systems, this parameter simply means that the first three letters of the first word and the first letter of each of the next three words (or as many words as there are) are used in the transaction code to abbreviate the title. Sometimes two or more titles answer to the same 3111 code. When this happens, the titles are listed out on the screen and the cataloger-recorder selects the correct one.

On the CRT screen, the record is practically identical to the old four-by-six title card. Title appears at the top, and issues are listed in the proper order below it as before. When data fills one screen, it may be continued on another, and it is just as possible to "page" backward and forward through them as it is to flip through cards in a file. In addition, the screen contains a few prompts, such as MESG for "clipping a note" to a record, XREF for claimant cross-references, and so on.

At the end of the catalog period, a proof is generated of the entire contents of the disc file and sent to an editor. At the same time, all titles (without their issues holdings) are transferred to a second disc file, which is now addressed by the cataloger-recorders in recording issues

for the new catalog period. Unlike the batch process used in making corrections to the proofs of the other catalogs, however, the periodicals corrections cycle is entirely on-line. The editor addresses the record in the now closed file and makes whatever additions, deletions, or modifications are necessary. When this task is completed, camera-ready copy is produced as usual, and the contents of the closed file are dumped into the archive file.

The advantages of the new periodicals system are fairly obvious: it is no longer necessary to sort periodicals by title, the duplication of effort expended in typing up the penciled cards is eliminated, and no longer is handwriting legibility a job requirement for a cataloger-recorder. On the other hand, year dates and registration numbers are far more susceptible to error when keyed than when penciled. As a result, greater care and more time must now be allowed for editing the work of the front-line staff.

THE FUTURE

Just as *standardization* and *compatibility* are of ever-increasing importance to international networking, so they have become the watchwords of the Cataloging Division's chief, Robert D. Stevens. Recent enactment of the first major revision of the copyright law since 1909 (October 19, 1976, P. L. 94–553, 90 Stat. 2541) will probably produce little in the way of fundamental change to COPICS. But it provides an opportunity to breathe new life into old concepts, and serious study is underway to explore the possibility of compatibility with other data bases. If the hurdles amounting to a basic difference in philosophy between a data base oriented toward bibliographic research and another toward the researching of a legal claim are cleared successfully, perhaps a base record prepared by one can be used as input for processing by the other. Other possibilities are under discussion: perhaps more attention can be given to establishing and using established name authorities after all; perhaps the full spectrum of diacritics and modified roman characters should be used; perhaps subject headings are possible; perhaps full MARC format is applicable; and so on. In any case, the Cataloging Division can at least accommodate ISBD format, which moves a step in the direction of compatibility even if no other accords are reached. In view of the wider variety of format and greater dissemination of information that COPICS now makes possible, whatever course is followed promises to place copyright records into the hands of more libraries, law offices, arts and sciences industries, and, perhaps, homes than ever before.

A Model of the NELINET Computerized Interlibrary Loan System: Testing Strategies for Load-Leveling

James WOLPER and Libby TRUDELL: NELINET, Wellesley, Massachusetts.

A mathematical model of a regional interlibrary loan system was developed as a tool for the design and evaluation of a strategy to equalize the load of ILL on libraries in the system. The model simulates an automated ILL system that keeps track of each library's ILL activity and identifies those libraries that are overburdened with requests. The load-leveling capability of several algorithms for representing each library's ILL activity was tested. It was found that the algorithms were nearly equal in effectiveness but presented differing implementation problems. Because of recognized limitations in the environment represented to the model, further evaluation of the load-leveling concept in a test version of an automated ILL system is suggested.

INTRODUCTION

From September 1976 to August 1977, the New England Library Information Network (NELINET) conducted a study on the effects of incentives for load-leveling on resource sharing. This research, based on the development of a computerized interlibrary communications system, was supported by a grant to the New England Board of Higher Education from the United States Office of Education Library Research and Demonstration Program.

The ultimate goal of the study was to design a computerized regional ILL system that would provide borrowing libraries with information to help them select the recipients of their ILL requests. One special aim was to devise a strategy for load-leveling that would prevent lending libraries from becoming overburdened with requests. This load-leveling

Manuscript received November 1977; revised February 1978; accepted April 1978.

concept was founded on the widely held assumption that large research libraries receive a disproportionate percentage of ILL requests when compared to the amount they borrow.[1] It was hypothesized that these overburdened institutions could be identified by monitoring requests made through an automated ILL system. If libraries using the ILL system to get location information for ILL purposes would agree to avoid borrowing from these high impact institutions (except in cases where no other holding library could be identified), the burden of ILL on them should be reduced.

In the proposed automated ILL system, the following information would be provided for each request: a list of libraries that own the item, obtained from OCLC and other data bases; fees, if any, charged by each institution; turnaround time at each institution; and a load-leveling incentive factor called, for our purposes, the institution's "market value." The system will identify those institutions from which borrowing should be encouraged and assign them a high market value; it will assign a low market value to those institutions from which borrowing should be discouraged.

Rather than designing a load-leveling scheme and testing it in an operational automated ILL system, the NELINET project staff decided to design a mathematical model of a regional ILL system as a test bed for a number of approaches to market value strategies. The model would enable the staff to examine several different strategies for assigning market value and to evaluate their performance in a variety of environments. In this way, a workable strategy for load-leveling could be chosen for implementation in the proposed computerized regional ILL system.

While previous studies of mathematical modeling of ILL have been made,[2] none has addressed the problem of load-leveling. The purpose of this article is to introduce the model and describe how it was used in selecting a market value strategy for the proposed NELINET ILL system.

THE MODEL

The model was designed and implemented as a FORTRAN program on the NELINET PDP/11. Its central function is the simulation of ILL transactions. The environment represented by the model is a set of libraries that engage in ILL traffic among themselves. Each library is an abstracted entity defined by the two most important characteristics of a participant in an ILL system: the borrowing demand it makes on the system and its ability to supply requested materials.

The load-leveling concept requires the introduction of a variable factor assigned to each library. This factor, called "market value," reflects the library's recent ILL activity. The definition of market value is based on the following axiom: libraries that have been disproportionately heavy borrowers should be assigned a high market value, while dispro-

portionately heavy lenders should be assigned a low market value.[3] In the model, libraries with high market value generally receive ILL requests in preference to those with low market values. The object of the model is to determine whether the assignment of a market value factor can effect an equalization in the ILL burdens of the libraries in the system.

The model performs three tasks for each simulated ILL transaction:

1. Designates the requesting library:
   For the purposes of the model, the requesting library is picked randomly. The probability of a library being picked is based on its assumed borrowing demand.
2. Determines which libraries hold the desired item:
   The model must determine how many libraries hold the item. This is done using a pseudo-random-number generator that reproduces the distribution of holders per item in the OCLC data base, as determined from the OCLC Cataloging File Statistics. Then, a corresponding number of libraries are designated as holders. The probability of any library appearing on this list of holders is based on its assumed ability to supply requested items.
3. Chooses the library that receives the request:
   When simulating a situation in which there is complete compliance with load-leveling, the system chooses the holding library with the highest market value. However, since it is not always desirable to assume complete compliance with the market value incentive strategy, a certain percentage of requests may be sent to any of the holding libraries without regard to their market values. Similarly, a certain percentage of requests may be sent completely randomly, in order to simulate the occurrence of blind searching.

Thus, a library's ability to supply requested items does not change; that is, its probability of appearing on the list of holders remains the same. However, the market value strategy will cause the demand made upon those resources to change, thus influencing its likelihood of being chosen from the list of holders.

APPLICATION OF THE MODEL TO THE PROPOSED NELINET SYSTEM

To test a market value strategy's effect on a given library environment, that environment must be represented for the model. Specifically, it must be decided how to represent to the computer a library's demand on the system and its ability to supply requested items.

The model environment used by NELINET is based on statistics gathered by the NELINET ILL Base-Line Statistical Survey of selected libraries in New England. Attributes of ninety libraries from those surveyed were assigned to libraries in the model.

Each library's demand on the system is defined so that it borrows the

same number of items in each simulation month as it borrowed in the month of the survey. Thus, borrowing becomes a constant in the model, while the lender of the item may change.

Several methods for representing each library's ability to supply requests were considered. One possible definition would have been to base it on library size. This could be rationalized by inferring that the more items a library owns, the more likely it is to hold any particular item. But size is difficult to measure (should it include bound journals, manuscripts, microfilm, etc.?). Also, not every item a library owns is available for interlibrary loan. It was decided that size alone is not a good measure of a library's value as an ILL resource.

Therefore, for NELINET's purposes, this factor is based on the number of loans reported by each library during the survey. This definition does not mean that the number of loans by each library is constant, since a library's market value affects demand upon it.

The strategy chosen has certain effects. It tends to make very large libraries appear overly valuable, since they received a disproportionately large number of ILL requests during the survey. Libraries that did little or no lending during the survey may be more valuable than they appear, since the lack of lending may have been due to poor availability of holdings information or causes other than lack of resources. Moreover, it preserves the effects of fees: those who charged for ILL services did less lending and so appear less valuable. A striking example of this is the Harvard College Library ($8 fee), which lent half as many items as Massachusetts Institute of Technology (no fee).

The effect of these tendencies is to make it more difficult for any market value strategy to foster load-leveling. Thus, this environment provides a rigorous test of a market value strategy's load-leveling capabilities.

MARKET VALUE STRATEGIES

There are two essentially different approaches to the computation of market value: continual and batched. The basic idea behind both approaches is to look at recent activity, judge how far from reciprocity the library's behavior has been, and adjust its market value to either encourage or discourage borrowing from that institution.

Continual adjustment involves an adjustment of a library's market value after each transaction: increasing a library's market value each time it borrows and decreasing it after each loan. Variation in this method results from the way in which the amount of increase and decrease is determined. It may be a constant figure, or it may be based on the library's activity.

The batched approach leaves a library's market value unchanged for some specified amount of time (or number of transactions) and then adjusts it based on its recent activity. As an example, assume each library's

market value was being adjusted after it had been involved in 100 transactions, and that in one such period some library had lent 61 items and borrowed 39. The excess lending indicates that this library's market value for the next transaction period should be lower, making it less attractive as a lender. Variations of the batched method fall into two classes: those that involve the assignment of the adjustment factor and those that involve the time between recomputations ("transaction period"). The transaction period may be determined by the number of transactions in which a library has participated or by a number of transactions in the system as a whole.

NELINET tested the following implementations of these strategies using the model:

Strategy 1: After a library has been involved in 100 ILL transactions, assign it a market value of $[(2M) + (B - L)] \div 3$, where M is the previous market value, B is the number of borrowings in those transactions, and L is the number of loans.

Strategy 2a: After each transaction, add 1 to the market value of the borrower and subtract 1 from the market value of the lender.

Strategy 2b: After each transaction, add to the market value of the borrowing library a factor inversely proportional to the probability of that library borrowing an item; that is,

$MV_{new} = MV_{old} + (K/P_B)$, where

MV_{new} = new market value

MV_{old} = old market value

K = a constant

P_B = probability of borrowing.

Conversely, a factor inversely proportional to the lending institution's probability of holding any item is subtracted from its market value; that is,

$MV_{new} = MV_{old} - (K'/P_L)$, where

K' = a constant

P_L = probability of lending.

EFFECTIVENESS OF STRATEGIES

Two criteria were used for measuring the effect of a market value strategy, one based on the reciprocity achieved for individual library performance and one based on the impact on reciprocity in the system as a whole.

Individual Library Reciprocity

For an individual library, the amount of reciprocity can be gauged by looking at the ratio of loans to borrowings. Ideally, this should be one to

Table 1. Ratio of Loans to Borrowings, by Library

Library	*No Strategy*	*Strategy 1*	*Strategy 2a*	*Strategy 2b*
A	1.95	1.24	1.39	1.36
B	0.10	0.68	0.21	0.58
C	2.54	1.51	1.37	1.46
D	0.11	0.64	0.92	0.80
E	0.17	0.73	0.66	0.65
F	19.57	—	16.57	8.14

1 = Ideal reciprocity ($B = L$)
Less than 1 = Net borrower ($B > L$)
More than 1 = Net lender ($B < L$)

one, so that distance from one represents how far a library is from perfect reciprocity.

In table 1, the six libraries are representative of particular types of ILL behavior. There were remarkable similarities in the effects of these three strategies on the individual libraries. In all cases, a significant impact has been made. For example, Library C is a heavy net lender with a large and famous collection, which receives and fills many ILL requests. The ideal would be for it to receive requests only when it is the sole holder of that desired item. This happened with Strategy 2a and nearly happened with the others. Library D is a small library with little ILL volume but a tendency to borrow more than it lends. Strategy 2a brought it very close to 1 (.92).

Systemwide Reciprocity

Systemwide reciprocity can be measured by the average of the absolute value of the difference between borrowings and loans for each library. This is ideally zero. The difference of the factor from zero measures the inequality between borrowing and loaning that exists in the ILL system as a whole.

Very similar systemwide reciprocity performances were achieved by the three strategies, as shown in table 2. Strategy 2b achieved the lowest systemwide imbalance by a very small percentage. None of the strategies brought the imbalance close to zero.

Evidently, the minimum value of the systemwide reciprocity factor is about twenty for this set of libraries. This is probably due in part to the large number of unique items shown by the OCLC Cataloging File Statistics, which were used to define the number of libraries likely to

Table 2. Average of Difference Between Borrowings and Loans, Systemwide

No Strategy	*Strategy 1*	*Strategy 2a*	*Strategy 2b*
35.00	21.98	20.60	20.22

hold a given item. According to OCLC statistics, approximately 15 percent of the items in the data base are held by only one participating library. No control can be exercised over these items. Another cause is the inequality in the library collections represented in the model; that is, that some libraries have a much stronger propensity to loan than other libraries.

Limitation of the Model Environment

The measurements in table 2 of systemwide reciprocity were obtained in an environment that assumed the holding library with the highest market value would always get the request. However, in a real library environment, this would be unlikely to happen because of the many factors that could affect a library's choice of potential lender. Therefore, these strategies were tried in several environments, wherein random holding libraries received requests from 5 to 25 percent of the time rather than all requests going to the holding library with highest market value. At 25 percent, severe degeneration of the effectiveness of the market value strategy was seen, as noted in table 3. When the amount of randomness in the system reached 25 percent, very little load-leveling was accomplished for any of the sample libraries. In fact, the ratio of loans to borrowings was further from one in some cases than it was with no load-leveling strategy.

In terms of systemwide effects, a linear dependence between the amount of randomness in the system and the degradation of reciprocity was found as randomness was increased, as shown in figure 1. At 25 percent randomness added into the system, the average of the absolute value of the difference between borrowings and loans for each library rises to forty.

The results show that load-leveling in this environment can occur only when the library with the highest market value is chosen more than 75 percent of the time. It is difficult to tell whether this environment exists in a regional ILL system. There are no data available on how many times librarians will ignore the market value; perhaps 25 percent of the time is too high a guess. Also, it is likely that the character of the items

Table 3. Ratio of Loans to Borrowings, by Library (25 Percent of Loans by Random Holding Library)

Library	*No Strategy*	*Strategy 1*	*Strategy 2a*	*Strategy 2b*
A	1.95	1.75	2.04	2.03
B	0.10	0.14	0.13	0.13
C	2.54	2.18	1.96	1.94
D	0.11	0.17	0.17	0.16
E	0.17	0.24	0.24	0.24
F	19.57	17.28	21.71	14.69

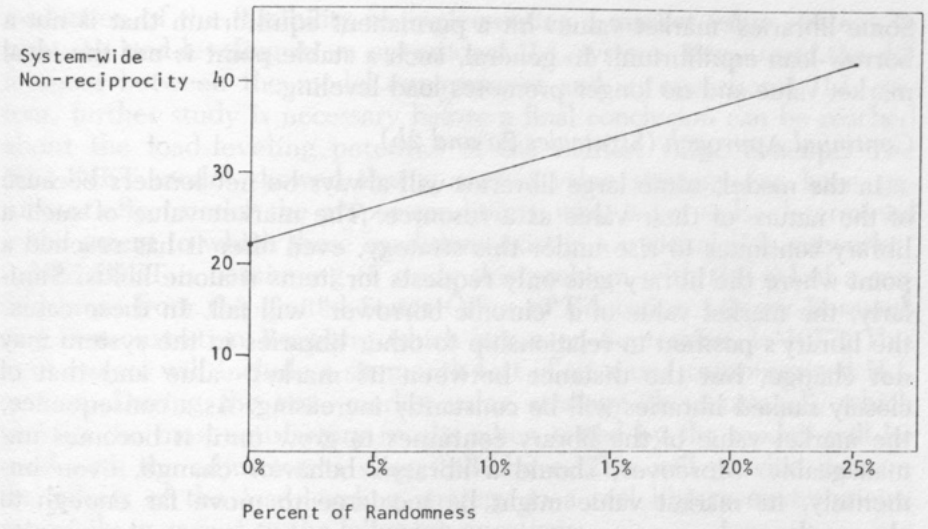


Fig. 1. Effect of Random Factor on Systemwide Nonreciprocity.

requested via ILL differs substantially from the character of the OCLC data base. It is possible that the number of items with only one holder requested via ILL is much smaller than their occurrence in the data base. This hypothesis is supported when one realizes that many ILL requests are for an item that has been cited, which makes that item a more attractive addition to many collections. Further, the OCLC data base does not accurately represent holdings of serials, and serials form a large part of ILL traffic. Thus, although more than 25 percent randomness is an insupportable amount in the model environment, it may not be insupportable in a real library environment.

EVALUATION

The similarity of the results for the three strategies indicates that the choice of market value strategy should be based on factors that affect ease of implementation in an ILL system as discussed below, since their load-leveling capabilities are approximately equal.

Batched Approach (Strategy 1)

The batched approach presents two problems as used in this model. For libraries that lend as many items as they borrow, this approach causes market value to shrink in absolute value; that is, all market values tend to move toward zero. It would be preferable to have a strategy in which such libraries maintain a constant market value; then, these can be used as a yardstick to measure how far market values of libraries not in reciprocity should move.

The other problem with this strategy is the existence of stable points.

Some libraries' market values hit a permanent equilibrium that is not a borrow-loan equilibrium. In general, such a stable point is not the ideal market value and no longer promotes load-leveling.

Continual Approach (Strategies 2a and 2b)

In the model, some large libraries will always be net lenders because of the nature of their value as a resource. The market value of such a library continues to rise under this strategy, even after it has reached a point where the library gets only requests for items it alone holds. Similarly, the market value of a "chronic borrower" will fall. In these cases, the library's position in relationship to other libraries in the system may not change, but the distance between its market value and that of closely ranked libraries will be constantly increasing. As a consequence, the market value of the library continues to grow until it becomes unmanageable. Moreover, should a library's behavior change, even immensely, its market value might be too large to move far enough to change the rank.

However, it is possible to change the adjustment factors of such libraries so that their market values would stabilize yet could still respond to changes in behavior. This could be done by manipulating the ratio between the factor added when a library borrows an item and the factor subtracted when it makes a loan, e.g., if the library consistently lent three items for every two it borrowed, market value would be stabilized by adding 3/2 for every borrowing and subtracting 1 for each loan. Another approach would be to stabilize the market values of these libraries at the point where they start to lend only unique items, since no further improvement in load-leveling can occur at this point. The difficulty is in deciding which libraries need this treatment and how often to revise it. An even simpler solution would be for the system to monitor a list of high impact libraries and redirect requests to another library unless the high impact library is identified as the unique holder of the item.

CONCLUSIONS

The primary intent of the mathematical model of the regional ILL system was to aid the development of a workable market value strategy that could be implemented as part of NELINET's proposed automated regional ILL system. The model provided a useful context for preliminary research to explore different approaches to load-leveling and enabled the project staff to develop and compare a number of different strategies in a simulated environment. Such a "debugging" phase would have been extremely difficult to carry out in the environment of an operating ILL system. Thus, the model served as a very valuable research tool in this project.

However, the model is limited in usefulness in terms of the final

evaluation of the feasibility of implementing a market value strategy to promote load-leveling in an operational ILL system. Because of the differences between the model environment and an operational ILL system, further study is necessary before a final conclusion can be reached about the load-leveling potential of the market value concept. The NELINET model showed that a market value strategy can have significant effect under the proper conditions; what is yet to be determined is the extent to which these conditions exist in a regional ILL network.

NELINET is continuing to study this problem with the aid of a second grant from the United States Office of Education Library Research and Demonstration Program, which began on September 1, 1977. This new project will include a six-month test of an actual computerized ILL system. During the test, market value strategy 2b (continual), which yielded the most satisfactory results when tested by the model, will be used with the adjustments suggested above. This will provide data on the effects of the load-leveling strategy in a real library environment, especially in regard to the following questions:

- How often will the library with the highest market value be selected by users to receive ILL requests?
- How many items requested for ILL are unique within the region, as measured by the number of holders indicated by OCLC?
- How many items requested are not cataloged on OCLC?

Definitive answers to these questions will help NELINET decide whether the proper conditions for a market value strategy as part of an automated ILL system exist in New England and can be used to help explore ways to make these conditions exist if they do not.

REFERENCES

1. Palmour, Vernon, and others, *A Study of the Characteristics, Costs, and Magnitude of Interlibrary Loans and Academic Libraries* (Westport, Conn.: Greenwood, 1972).
2. Rouse, William, and others, *A Mathematical Model of the Illinois Interlibrary Loan Network Project Reports, Nos. 1–6* (Urbana, Ill.: Coordinated Science Laboratory, 1974–76).
3. The measurement of interlibrary loans is a matter of some controversy, and the term "loans" must be used with care. The effort involved in processing any ILL request, whether it is filled or not, is considerable, and many ILL librarians feel that it is enough that *all* requests should count as loans. Others feel that an ILL system is only concerned with the transfer of materials, so only actual loans should count. Eventually, a policy decision on this issue will be made by the NELINET Committee on Interlibrary Communication.

Statement to the House Communications Subcommittee

Important deliberations are now going on in the House Communications Subcommittee regarding the proposed revision of the Communications Act of 1934. Subcommittee Chairman Lionel Van Deerlin (D-Calif.) introduced a revision bill, HR 13015, on June 8, 1978. A new law, once enacted, is likely to affect many aspects of telecommunications use in the United States for the remainder of this century and well into the next. The following statement establishes the vital interest libraries have in this matter and presents many of the issues confronting the library community with respect to the formulation of an enlightened communications policy. This statement was prepared by the Washington Office of the American Library Association with assistance and contributions from several members of the LITA division (then ISAD) who are acknowledged in the concluding paragraph.

STATEMENT OF THE AMERICAN LIBRARY ASSOCIATION
TO THE COMMUNICATIONS SUBCOMMITTEE
OF THE COMMITTEE ON INTERSTATE AND FOREIGN COMMERCE
U.S. HOUSE OF REPRESENTATIVES
ON REVISION OF THE COMMUNICATIONS ACT OF 1934
NOVEMBER 23, 1977

Edited by Carol C. Henderson
ALA Washington Office

INTRODUCTION

The American Library Association, a nonprofit educational association of some 35,000 libraries, librarians and information specialists, library trustees, educators, and communicators, fully supports the efforts of the House Communications Subcommittee to revise the Communications Act of 1934 to accommodate new technologies. Methods of communication have changed radically and the volume of communications has increased many times over since the 1934 act was formulated. It is estimated that over half the U.S. work force is now engaged in the processing of information. In the midst of this information revolution, the American Library Association is particularly concerned that the public's access to meaningful and coherent information not be lost. The subcommittee is in a crucial position to insure that the average citizen's right to information includes the right to the kinds of information and the modes of access to information that the new technologies will provide.

At the suggestion of the subcommittee staff, this statement has been prepared in response to the staff's "Option Papers," and we therefore ask that it be included in the subcommittee's hearing record. It is organized according to the topics of those papers. The subjects covered are public broadcasting, domestic communications common carrier policy, international telecommunications, options for cable television regulation, and the impact of communications technology on the right to privacy. As noted at the conclusion of this statement, the various sections have been prepared by members of the American Library Association and its Information Science and Automation Division and by ALA staff.

PUBLIC BROADCASTING

The American Library Association feels that the House Communications Subcommittee's staff option paper on public broadcasting has rendered a valuable service by delineating current and emerging issues in this field. We are pleased to note that the president's set of proposals (House Doc. 95–239) sent to Congress recently contain some of the ideas found in the option paper and is for the most part a comprehensive and reasonable proposal. We wish to make the following comments about both the option paper and the president's proposal.

I. Nonbroadcast Technology

We applaud the intent of the subcommittee to augment and expand the basic system of public broadcasting by the use of other distribution techniques in the nonbroadcast technology area. A public telecommunications system which would utilize all available technologies could provide many more program choices for the public through multiple outlets, more adequately meet the needs of many separate or different audience groups, and encourage new creative talent both at the local and national levels. We feel the focus has too long been on *broadcasting* to the neglect of valuable *nonbroadcast* educational and informational resources, which remain untapped and which are more easily responsive to local and regional needs. We would, however, extend this concept still further to include print as well as nonprint technologies. A comprehensive system of resources and media would seek to integrate resources of all types—broadcast, nonbroadcast, print, film—into a cohesive whole, with each being used to do the things it can do best and to reach the widest possible audience and serve the greatest number of people. In short, a pluralistic society demands a pluralistic communications system.

II. Educational Broadcast Facilities Program

We express strong concern, however, about the transfer of the Educational Broadcast Facilities Program from the Department of Health, Education and Welfare to the Corporation for Public Broadcasting. We understand the advantages of concentrating systems planning in one agency, but we feel the equipment and technology dollars should be separated from the programming dollars. It is our recommendation that this program be left at HEW and given a much higher status than it now has within that body. Coordination and cooperation between HEW and the programming entities (CPB, PBS, NPR) are absolutely essential but consolidation of the two is undesirable and risky.

If insulation of programming from pressures and from improper political influence is a desired goal, then consolidation does not achieve this goal. Such a move would invite politicization from powerful groups within the public broadcasting industry. In addition, too much power would be vested in one agency. A system of checks and balances is needed. By moving the facilities program to CPB, any effective liaison by the Congress itself with the program would be ruled out. Insulation from improper political influence in programming is certainly desirable, but insulation in the case of the facilities program would be unwise.

The White House proposal would turn federal dollars over to a nonfederal private agency with no stipulation or guidelines as to where, when, and how much money would be spent—and for what purposes.

Nonbroadcast technologies would, in our judgment, receive short shrift at CPB, where staff is oriented primarily to broadcasting. The system would continue to be centered around broadcast technology, with all other technologies being forced into a supplementary and secondary role. We feel that groups other than public broadcasting stations should be permitted and encouraged to apply for licenses to operate an Instructional Television Fixed Service (ITFS) or cable system (e.g., public libraries, school districts, other public groups).

In the event that a decision is made by Congress to move the Educational Broadcast Facilities Program from HEW to CPB, it is essential that a line item for facilities be included in the CPB budget, with specific amounts mandated for the facilities program. If this is not done, there is a danger that the money will be added to the general CPB budget and used for other purposes. We also recommend that separate line items appear within the facilities portion of the budget for broadcast and for nonbroadcast technologies if a balanced program of telecommunications is to be achieved. We urge that two-thirds of the proposed budget authorization ($20 million) be allocated to broadcast facilities, and that one-third ($10 million) be allocated to nonbroadcast facilities.

The facilities program is now a part of the Office of Libraries and Learning Resources at the U.S. Office of Education. Educational institutions throughout the nation have combined print and nonprint resources in the same administrative unit so that libraries are now typically instructional media/resource centers. Librarians and media specialists are trained in both fields. Radio/television activities in many school districts are an integral part of the educational program and are located within the resource center unit. Moving the facilities program to CPB tends to destroy the marriage between print and nonprint resources. One possible solution to this problem might be to move the Educational Broadcast Facilities Program to CPB, while keeping the nonbroadcast technologies at HEW so that this coordination with libraries and media centers would be preserved.

III. Public Participation

The American Library Association is deeply concerned at the lack of any mention in the White House bill of the importance of public participation. We feel the time has come to put the public back into public broadcasting. CPB's track record in the area of public involvement in its activities leaves a great deal to be desired. The recent and disappointing dissolution of the Advisory Council of National Organizations to CPB leads us to have serious reservations regarding the corporation's ability to enlist the public in any meaningful way in the decision-making process of CPB or in its programming efforts. The Yankelovich study, *Public Participation in Public Broadcasting*, released earlier this year, points out the problem very succinctly:

> The Board's principal obstacle is not its failure to accept the principle and theory of public participation, but rather, the fact that they have not thought through how to implement this principle in a practical and effective fashion. It is perhaps not an exaggeration to suggest that the Board, while committed to public participation in principle, in practice has not specified what decisions it applies to, how to mobilize it, what tradeoffs and inconveniences to accept in incorporating it into their decision-making process, and how to benefit from what it has to offer.
>
> To the extent that this observation is valid, the undeniable problems of working with the existing ACNO structure cannot be fruitfully addressed until the Board first resolves these issues. As matters stand today, even if the Board were given a totally trouble-free mechanism for achieving public participation—such as an imaginary computer that would automatically answer any questions about public attitudes without demanding anything in return—it would, under existing conditions, probably not be used very often.[1]

We urge the House Communications Subcommittee to include in its version of the Public Broadcasting Financing Act of 1978 a strong statement regarding the

urgent need for CPB to give priority attention to the development of new mechanisms that would improve public participation. At the time of its dissolution last September, ACNO urged the CPB board to devise "new methods and new structures to answer the many challenges now facing not only CPB, but public broadcasting as a whole," and also urged the board to appoint a task force to develop proposals for insuring future public participation. The board did not act on this until November 11, and then only authorized its Public Affairs Committee to set up such a structure. It is likely that such a task force will not be created for several more months, and during this period CPB will have virtually no public participation mechanism in operation. We urge Congress to use its influence to persuade CPB to accelerate the timetable for this task force's work and to require CPB to set aside a percentage of its budget for the creation of a public participation mechanism that would be independent of the corporation and insulated from internal industry pressures. This participation mechanism should require CPB to seek input from educational organizations as well as public broadcast station personnel.

IV. Programming Responsive to Local Needs

There is an urgent need to restore localism to a system that has become increasingly national in scope. Only 11 percent of the programs aired on public broadcasting are produced locally. This percentage is far too small. There is striking similarity between public libraries and public community broadcasting stations. Both are publicly supported informational, cultural, and educational institutions. There should be and in some communities there is collaboration between the two in the coordination of services and the promotion of those services. Public libraries are of necessity responsive to the needs of their local communities, and public broadcasting stations should be to a much greater extent than now exists. One method of doing this would be to extend the ascertainment process, which requires commercial stations to determine the programming preferences of their audiences, to public broadcasting stations.

To extend the reach of public broadcasting programming, in localities where cable television is operative, libraries would like to acquire programs developed by PBS so the public could have access to them on cable television or in-house through playback videotape machines. Unquestionably, diversified technology can provide mechanisms which will make possible multiple channels for reaching the public with a diversity of programming. Every effort should be made to use these technologies to extend the basic service which public broadcasting provides. It is becoming increasingly more expensive to provide local broadcast outlets in areas which are not already being adequately served. Additional technologies must be tapped and coordinated with public broadcasting stations if the maximum possible audience is to be reached. Nonbroadcast technologies can open up a broader range of services than program transmission can provide—services which are essential to the transmission of data, increased access to information, and future two-way interactive uses.

V. Eligibility for Grants for Public Telecommunications Facilities

Finally, ALA applauds the inclusion in the White House proposal under Section 392, "Grants for Construction and Planning," of the provision that applicants for grants can be *other than* public broadcasting stations. This provision has long been needed. We highly endorse the inclusion of the phrase under 392(a) (1) (C) of "a nonprofit foundation, corporation, institution, or association organized primarily for educational purposes" as one possible grantee under the grant program for construction and planning of public telecommunications facilities. This will enable public libraries, school districts, and educational institutions to be eligible for grants without having to go through the public broadcast station in the community. A public telecommunications center, as envisioned in the option paper, is an innovative concept which needs to be developed in each community, but it does

not always need to be operated, controlled, and housed at a public broadcast station. Likewise, a public library might wish to be the hub of a community-owned public cable system, or a school district might wish to own and operate an ITFS facility. This flexibility should be possible under the provision of Section 392, and we urge its retention.

In short, we urge the subcommittee to develop flexible and innovative public broadcasting legislation that will increase the options of all Americans to obtain in their communities diverse and high-quality programming. In working toward this goal, we request that you keep in mind the role of library and information services, which have much to offer public broadcasting.

REFERENCE

1. *Public Participation in Public Broadcasting:* The Findings of a Study by Yankelovich, Skelly and White, Inc. (April 1977), p.9.

DOMESTIC COMMUNICATIONS COMMON CARRIER POLICY

Modeled after the 1887 Interstate Commerce Act at a time when communication by wire or radio seemed similar to transportation by rail or barge, the Communications Act of 1934 is by any standard an outmoded framework for communications common carrier regulation today. This statement discusses, first, the role of libraries in national communications policy; and second, suggested directions we believe a revised communications common carrier policy should take.

I. Role of Libraries in Communication

Librarians approach the subject of communications from the perspective of public access to information. We believe that all Americans, regardless of their geographical location, their work, or their financial resources, should be able to find quickly and easily the information they need in order to lead productive lives. Because the strength of the nation depends upon just such an informed citizenry, we believe equal access to information is a matter of public policy. In this connection, we call to your attention the Statement of Policy set forth by Section 2 of Public Law 91–345, the National Commission on Libraries and Information Science Act, which provides:

> Sec. 2. The Congress hereby affirms that library and information services adequate to meet the needs of the people of the United States are essential to achieve national goals and to utilize most effectively the Nation's educational resources and that the Federal Government will cooperate with State and local governments and public and private agencies in assuring optimum provision of such services.

Increasingly today access to information requires utilization of communications and computer technologies. No longer is information synonymous with the book. While the old words remain—"library" derived from the Latin word for book, "bibliography" from the Greek word for book—the information services provided by libraries today are likely to involve machine-readable files or data bases in an interactive on-line format. Such on-line bibliographic search services are generally tailored to the serious researcher, the government official, or persons in business and industry, but increasingly, as budgets permit, libraries are making machine-readable data bases part of their regular reference and information services. Interlibrary cooperation enables users of some of the smaller libraries to avail themselves of such services as well.

New technology has spurred the development of a new form of interlibrary cooperation—the computerized library network. Distance between libraries becomes insignificant by virtue of such a network, which allows a library to join with many others in programs of shared access to cataloging data or shared collection development. Economies of scale are thus effected, and duplication of effort is greatly reduced. As a result of such networking projects, local access is extended to the collections of other libraries. Networking thus has the potential of greatly expanding public access to information by making the library resources of the entire nation available to whatever individual may need them.

The library resources of the nation in-

clude the collections and services of: 8,500 public libraries and nearly 5,500 branch libraries; some 2,800 college and university libraries; 6,500 special libraries; some 2,000 medical libraries; nearly 1,000 law libraries; and more. Although libraries exist in communities throughout the country, their resources are not evenly distributed geographically. Cities like New York, Philadelphia, or Boston have prodigious library resources, while some of the rural states have relatively few. Library networking carries with it the potential to move information rapidly to individuals and groups rather than requiring people in need of information to travel across the state or to a neighboring state to find it.

In America, libraries have a long tradition of service to the public. For example, one need pay no tuition and no admission fee to visit a public library and use its varied collections. Local and state governments, as well as private philanthropy, have all contributed to the support of libraries. In 1956, the U.S. Congress added federal support to libraries by enacting legislation (the Library Services Act) to assist the states in extending public library service to rural areas. All Americans should have access to the knowledge stored in libraries, Congress concluded, not just those who live in the cities. This move on the part of the federal government in the 1950s to ensure the development of library service in rural areas parallels its actions in the two preceding decades to extend communications service through enactment of the Communications Act of 1934 and later the Rural Telephone Act. A policy of nationwide postal service with certain rates uniform also helped to equalize the communications and information services available to all Americans.

It should be clear by now that from our perspective any national communications policy is inextricably interrelated to national information policy. Access to the information resources of the nation is as important as access to communication services. New technological innovations have now made it possible for Americans to communicate in new ways. In 1934, the emphasis was on voice communication, but today digital data transmission is increasingly important. Libraries serve the vital function of bringing people together with the information they need. To accomplish this task libraries must be able to communicate not only person-to-person, but also computer-to-computer, terminal-to-computer, computer-to-terminal, by teletype, facsimile, and so forth. Nonprofit information agencies such as libraries perform a service that is clearly in the public interest. Whether or not they will be able to reach their full potential of assuring equal access to information for all Americans depends in large measure on whether and to what extent they are able to utilize the new information technology. To date, costs have raised the greatest barrier.

II. Common Carrier Policy Directions

To librarians, public service is the most important aspect of common carrier regulation. We would like to see a true consumers' regulatory agency and not one biased in favor of the industry under regulation. The regulators must be oriented toward consumer benefits, not industry profits. The original goal of the 1934 act is a fine one which we support today: making available to all the people of the United States a rapid, efficient nationwide and worldwide communications service with adequate facilities at reasonable charges. It seems likely to us that costs will be more "reasonable" with the authorization of customer-provided terminal equipment and with intercity private line competition, and we support the FCC's recent efforts to open up such areas to competition. We also think that strict prevention of cross-subsidization is essential to encourage true competition, thereby reducing costs to consumers such as libraries, as well as encouraging technological innovations which might well be slower to develop in a monopolistic situation.

We do not believe that complete deregulation is a viable choice. We fear that the result would be complete or substantial market dominance by a single firm (or a very few firms) which would work against the interests of consumers. In the absence of regulation by the FCC, antitrust alone would be insufficient to protect consumers. While it could have a beneficial effect

in the long run, by its very nature antitrust regulation comes late in the game. Customers are exploited, often for long periods of time. Prosecution is difficult and takes years. We think that a combination of administrative regulation and competition is essential to achieve lower prices and the development of innovative technology and services. Because regulators tend to be pro-industry rather than pro-consumer, congressional oversight of the regulators is essential. Congress has been far too inactive in this oversight function. It is difficult to determine just where regulation should give way to free and unregulated competition. The present pattern of determining whether a service is primarily "data processing" and therefore not subject to regulation, or whether it is primarily "communication" and therefore regulatable, seems to us to be unsatisfactory. We have some hopes that as a result of the second computer inquiry now underway, the FCC may develop a more workable approach.

We would not support a requirement that all communications services pay their own way, for clearly there are some which it is in the public interest to support. However, the current subsidization picture is mind-boggling and somehow must be unraveled. Some cost data show that local calls are subsidized by long-distance calls. Other data tend to show that the opposite may be true. Somehow the costs of providing different services must be determined by an objective party. Once this determination has been made, Congress can then decide which services it is in the public interest to provide. If such services are unable to pay their own way, some form of direct government support should be provided. The record of the industry in providing reliable cost data upon which such policy determinations can be made has been deplorable. An objective determination is essential.

Separation of terminal equipment from the "separations and settlement" process should help to rationalize pricing and improve competitive conditions for the consumers, including libraries. "Unbundling" is a natural corollary of this. It is essential, however, that cross-subsidization be carefully policed to prevent an industry from using monopoly power in other markets to cross-subsidize anticompetitive and predatory terminal equipment rates. Standards for interconnection of terminal equipment are of primary importance to the development of a national library network. In order to promote technological innovation and competition among equipment suppliers, it seems to us that standards should be established by the federal government (National Bureau of Standards) under advisement from industry. Libraries must deal with lengthy, formatted textual messages in many different languages in digital form, and to the extent that the attachment and operation of so-called specialized terminal equipment is impeded by regulation or by monopolistic decisions made on the basis of other markets with radically different information-handling requirements, we oppose it. We believe that further opening up the terminal equipment industry to competition will result in a variety of new services.

As providers of information to the public, libraries of all types are vitally concerned with communications. We believe that Congress should declare that the public interest requires the development of new communications services for the public at large, and that a new consumer-oriented communications policy is in order. While certain large-volume users of AT&T services now qualify for reduced telecommunications costs, we believe that other criteria besides sheer size should be considered in the determination of who is to receive lower rates. So long as such determinations are left up to industry, it seems highly unlikely that the consumer will benefit. The task which the subcommittee has set is awesome; we all recognize that. Our plea here is simple, inelegant, and difficult for you to implement. We believe that public policy requires provision of communications services in a variety of formats and at reasonable prices which the American people through their institutions, such as libraries, can afford. We further believe that public access to information is a significant national goal interrelated with national communications policy. We hope that the work of the sub-

committee will lead our nation ultimately in the direction of a flexible and innovative consumer-oriented national communications policy. If you can do it, those who have participated in your deliberations and those who share our view will be reassured that our government can be responsive to common need rather than corporate leverage.

INTERNATIONAL TELECOMMUNICATIONS

In its option paper the subcommittee staff draws attention to the opportunities presented by the convergence of computer and communications technology. This technical marriage implies an enormous new capability for distributing information in all forms—print, audio, digital, analog, and video—around the world. For users of libraries and information centers, it brings promise of integrated access to a network of networks, which will tie together most of the world's current literature and data. Neither the computer nor the communications satellite had been invented when the basic Communications Act was enacted in 1934, and we agree that basic revision of the act is needed.

The option paper also addresses the need to structure the U.S. international telecommunications industry so that it represents the long-range telecommunication requirements of the nation as a whole. International telecommunications is an intricate, highly complex field. Because it has to do with foreign policy and America's dealings with other sovereign states, it is politically complicated as well. While most foreign telecommunications services are owned or operated by their governments, the U.S. has a telecommunications industry consisting of a number of giant private firms. If the companies proceed independently, as in the past, unilateral action by one company might contradict international agreements made in the name of the U.S. Different technologies, such as satellite and terrestrial communications, have developed largely on separate grounds and along separate lines.

We would support the establishment of a more centralized point of international communications coordinating responsibility than is now the case for this country. Such a step is necessary to enable the U.S. to speak on telecommunications with a more integrated voice. It would also provide a more centralized point at which public interest groups could make their voices heard when questions of rates, services, and applications arise.

The third major point addressed by the option paper is the risk of locking the U.S. into international telecommunications agreements prematurely and constraining our later reactions to emerging telecommunications demands. This risk is real and is particularly relevant where libraries and information activities are concerned. The library and information community is undergoing a gradual but noticeable change from a loose confederation of independent units to a national network of functionally interdependent organizations. It is not at all clear yet what this trend implies as far as telecommunications is concerned, but we can make some reasonable estimates. First, the volume of library telecommunications traffic is likely to increase—nationally and internationally. Second, the traffic will be mixed—consisting of print and nonprint forms. Third, the library of the future will be in increasing direct, two-way communication with the home and office.

A growing number of libraries in this country cooperate through computerized library networks to provide more equal access for citizens to the unequally distributed materials in libraries and information centers. Through such networks, users gain access not only to the collections of other libraries but also to the information in numerous machine-readable data bases. Similar developments are taking place in other countries. The next step is to connect U.S. library networks to information networks abroad. As we become more and more of an information society, two-way, on-line access to information may become as commonplace and as necessary as postal and telephone communications are today.

All Americans should have access through libraries not just to traditional published materials but also to information in newer, nontraditional, and perhaps

even yet to be developed forms. For this reason provision of adequate telecommunications for libraries and information centers is a public responsibility. The U.S. should plan international telecommunications systems in a manner which will allow the broadest possible use of it for public education and information purposes. Nothing less will meet the original goal of the 1934 Communications Act: making available to all the people of the United States a rapid, efficient, nationwide and worldwide communication service with adequate facilities at reasonable charges.

OPTIONS FOR CABLE TELEVISION REGULATION

The following are the main points the American Library Association wishes to make in response to the option paper on cable television regulation:

I. Cable television has not grown under the present regulatory structure. This structure must therefore be revised so that the industry will have a fair chance. Regulation as a common carrier is suggested.

II. Cable television should be considered as one communications medium among many, all of which must be considered as a whole, with legislation and regulations being formulated accordingly.

III. Cable television can be, and has proved to be, a unique medium for communicating local matters, as opposed to broadcast television which is a mass medium. Any forthcoming planning must take this uniqueness into account as the law is rewritten and regulations are being formulated.

IV. Libraries and other information services can and should use cable to communicate their messages to local client groups. Details of such communication will be spelled out.

I. Cable Regulation

Up to this time, the FCC has been responsible for regulating cable television. The 1972 rules envisioned cable as a medium of abundance which would carry all commercial broadcasts in a given area, as well as have channels set aside for education, government, and public access. Other requirements were based on this assumption of abundance as well: channels were to be set aside for foreign-language programs and other specialized services. Up until this time, it can be seen that the promise has not been fulfilled. Cable has not grown as rapidly as many studies had forecast and the uses envisioned for it have not materialized in most places. Reasons for this may be found in the reluctance of the FCC to let cable compete in the free marketplace, to import any stations and programs which it was able to obtain, to simplify paperwork in compliance with FCC, state, and local regulations, and to compete with broadcast television in other ways. The FCC has consistently heard testimony from broadcasters and program owners and disregarded such testimony from the cable industry and public interest groups (this may be inferred from their many decisions since 1972, which were designed to protect broadcasters and inhibit public interest use of cable systems).

We, therefore, suggest that program ownership and activities be divorced from signal carriage and that cable be regulated as a common carrier, subject to rate regulation and obliged to carry programming from any source able to lease a channel. Such charges should be sufficiently low for public interest groups and not-for-profit enterprises and organizations as not to make the use of a given cable system impossible for such groups. Pay cable, importation of sports broadcasts and distant stations, data communications, public access programming, etc., should all be able to use cable systems. In order to do this, regulation should insure a sufficient number of channels and great enough technical sophistication among systems so that these are capable of responding to commercial and community needs. This means a set of minimal technical standards, which must be met by all but the smallest systems (two-way capability, ability to interconnect, split, and amplify given signals, interface computers). It also means strict rules against cream-skimming:

a given community must be wired up equitably, not as seems customary now, connecting wealthier areas first and some poorer areas not at all. This latter practice means that those who may need information and other services most may not be able to receive them for years to come.

Federal regulation of cable should assure the above; all other regulation should be left to state and local jurisdictions. However, a proviso should be contained in the federal regulations that no duplicate records be kept nor data furnished that are arrived at in a different way for any level of regulation. In other words, paperwork must be simplified so that one set of requirements may be formulated that would fulfill the needs of all three possible levels at which regulation may take place.

II. Cable as Only One among Many Communications Media

According to Daniel Bell and others,[1] we are now living in a society in which information generation and dissemination constitute the largest single sector of the U.S. economy. To move this information from place to place, a communications system is needed which consists of computers and telecommunications. It has become exceedingly difficult to separate computers and ancillary equipment from telecommunications links which carry computer messages over distance, as is witnessed by the pending second attempt before the FCC to inquire into computer/communications problems. It is also exceedingly difficult to separate information (the "goods") from the machinery (the "vehicle"), e.g., computers and telecommunications (the "highways"). All three topics must be treated as a whole if the U.S. economy is to take full advantage of the information generated throughout the country so that it may help to accelerate the growth of the economy. Policymakers at the highest level have recognized the information-communications-computer complex of problems as a macro-problem and have formulated plans to cope with it on a macro level.[2, 3]

Cable represents but one of the many telecommunications devices or systems which may be used to carry information in printed (graphic), audio, or visual form over distance. Until recently, this distance was usually limited to a single city. By interconnecting via satellite, cable has become a competitive medium with the telephone system, as well as with value-added networks and special-purpose carriers. As yet, cable's interconnections are not used much beyond the transporting of pay-cable programs, but there is no reason why the system's wideband capabilities cannot be used to carry information and data of all kinds. A bank in New York is presently using cable to interconnect two branches;[4] Reuter's is offering a computer-based news service via cable.[5] Both examples demonstrate that such services are feasible. However, as long as cable system owners must struggle to make ends meet by concentrating their marketing efforts in such highly profitable areas as pay cable because they are not allowed to compete for other types of communication services, such experiments as the ones described above will remain isolated.

While cable should, by legislation and ensuing regulation, be encouraged to venture forth into new information services, it is but one medium in the nation's emerging communications network. It must be interconnected with other technologies (e.g., MDS, translators, and the like) so as to serve scattered rural populations, for example. It also will be using optical fibers instead of coaxial cable in the near future, because this medium is less expensive than coaxial cable. It will compete directly with the telephone system, which also envisions using optical fibers for home connections, thus for the first time becoming competitive with cable in the wideband area. There is no doubt that there will be struggle among competing technologies for their share of the market, and any legislation or regulation being formulated should take this into account. Policymaking in the telecommunications area should be based on the laissez-faire concept of the free marketplace, rather than artificially structuring the network in such a way as to favor one system over another.

Some legitimate questions to be asked include:

If a mixed technology network is

created, what will travel over this network? Who will have access to it? How much control should there be in order to insure privacy and confidentiality while preserving the public's right to know?

Will the traditional agencies (libraries, etc.), which have provided many of the information services in the past, be able to utilize the network or will it become so expensive as to prohibit use by all but private industry?

Who will insure that mixed technology networks will be interconnected in such a way that all parts of the country, and everyone, urban and rural, can be connected if need be?

How can we insure that people who need information may have access to it even though they are unable to pay for it (e.g., social service information in the cities)? Should there be a system of information stamps in the same manner as our system of food stamps? If this question is not answered satisfactorily, the "have's" (information elite) will have more, and the "have not's" (information poor) will have less.

It may very well be necessary that a policy decision at the highest level be stated to insure that information is a "public good;" that everyone has a right to access to this public good; and that the federal government must insure, through its many existing policies operating within this new conceptual framework, that the right to information is guaranteed. The locus of information/communications policy implementation should also be at the highest level; it is extremely doubtful that the position of the proposed assistant secretary for communications and information carries with it sufficient power to accomplish the extensive coordination of existing policies and the formulation and passage of new policies.

III. Cable as a Unique Medium

Because of the abundance of channels which a cable system can carry (with appropriate technology, about forty-four or even more), the inexpensive equipment needed to produce programs ($2,000 for minimal equipment, as compared to $250,000 for a minimal broadcast studio), and the ease with which the use of such equipment may be learned, cable is a medium which can be used by anyone with a message. No longer is it necessary to carefully script, rehearse, and stage a "show" in a studio; instead, members of a given community* can take a portable pack consisting of a TV camera and a video recorder anywhere, use existing light, and record an event, an instructional program, a zoning hearing or City Council at work. Those persons interested directly in the subject which has been recorded will watch a cablecast of it—in the same manner in which persons subscribe to the many highly specialized magazines, or belong to very specialized clubs and organizations—all to learn more about that in which they have a vital interest.

Libraries have recognized the value of cable for some time now. Over 200 produce their own video programs, others obtain them from elsewhere.[6] About seventy-five use the local cable system to disseminate their programs. The fact sheet attached to this statement provides details about some of these video/cable activities in libraries. Examples range from typical library programming, such as story hours for children, discussion groups, and instructional broadcasts, to showing local ballet and musicals, as well as sports programs, educational films, and speakers on particularly interesting subjects (a local "Meet the Press"), etc. Creative video artists have a chance to have their work displayed, and persons from widely differing political groups may be heard. An exchange of information concerning local matters is thus created; the community becomes much more aware of local concerns, and a forum for information exchange is created. That this is needed in this society which exhibits more and more symptoms of alienation and lack of concern for others goes without saying. That cable is the only medium to help in efforts to

*"Community" can be a geographic area or a community of like interests over distance, e.g., tennis players, balletomanes, aficionados of field trials, mushroom hunters—or parents of a handicapped child—the list is endless.

create a community spirit visually and aurally is also evident. The most competitive medium, the local newspaper, requires an ability to read—cable does not. Thus the 21 million Americans who are functionally illiterate can be brought into the mainstream of a given community with a minimum of expense and effort.

Beyond the applications outlined above, cable is capable of being used as a substitute for transportation in cities. Rather than bringing together persons from outlying branches for staff meetings and in-service training, such programs can be disseminated by cable; answer-back can be by telephone now, and the cable can be used wherever two-way capability exists. Printed materials can be displayed rather than shipped back and forth between agencies, saving wear and tear on the materials, and making them available to all simultaneously, rather than being out on loan. Many other applications are possible, involving computers which are being used in libraries to help control the ever-expanding numbers and types of publications, nonprint media, and the like. In order to use cable in this way, the minimal standards outlined above are mandatory. Otherwise, systems will be incompatible and unable to be used for data communication and two-way conferencing.

IV. Cable as a Social Service Information System

Besides the applications outlined above, there is one which needs particular emphasis. This is social service information dissemination by library-based *I*nformation and *R*eferral Services. An I&R service collects extensive data on human services available to all persons needing such service within a given area (usually a city) and disseminates that information via telephone to persons who have been made aware of the service by various means. Careful interviews are needed to elicit the exact problem and a good knowledge of the resource file to fulfill the need via referral. Often there is a three-way telephone which allows the I&R professional to insure that client and service provider are able to communicate. Follow-ups are often made to insure that quality service is obtained.

Because many poor people do not have telephones (while almost all do have television sets), the TV set becomes an ideal instrument to disseminate I&R information, provided there is a cable system which can distribute it via channels especially set aside for I&R services. At the most basic level, information can be typed on cards which are mounted on a so-called "message wheel" (now used for commercials) which is continuously spun while being scanned by a camera. At its best, such information is computerized and the computer can be programmed to display specialized services for specialized needs in response to inquiries, either by telephone or via terminals located in the viewer's home. Because of the increasing complexity of city and other human services, it has become necessary to establish I&R services everywhere; they are not well known and are usually underutilized. To promote them via cable, the local medium as outlined above, would be both inexpensive and beneficial to the many people needing this type of service. Any forthcoming legislation should take into account the need for the wide dissemination of social service information via cable and should encourage by all means possible such programming.

REFERENCES

1. Bell, Daniel, "Toward a National Policy in Information," address at 1975 annual conference of the American Society for Information Science.
2. Domestic Council. Committee on the Rights of Privacy, *National Information Policy; Report to the President of the United States* (Washington, D.C.: NCLIS, 1976).
3. "An Aspen Institute on Communications Reorganization," *Broadcasting*, p.26, 29+ (12 Sept. 1977).
4. "Info-Cable in New York Exciting Prospect for Librarians," *CableLibraries* 4:6–7 (April 1976).
5. "Reuter in Manhattan," *CableLibraries* 4:7 (May 1976).
6. "Results of VCCS Video Survey Presented in Committee at ALA," *CableLibraries* 5:1–2 (July 1977).

ATTACHMENT

Public Libraries Engaged in Video/Cable Activities

A recent Video Resources Survey conducted by the Video and Cable Communications Section of ALA's Information Science and Automation Division found that over 200 libraries produced their own video programs, and 75 libraries used their local cable system to disseminate their programs. The following examples of video/cable activities are just some of the ways in which libraries are changing their traditional roles and seeking new avenues of communication in responding to citizen needs.

Arizona

The Tucson Public Library initiated video services more than four years ago. The library has worked with community organizations and the school system to produce tapes of particular interest to area residents. Included are video programs in Spanish for the Mexican-Americans and programming for the elderly. Since November 1976, the University of Arizona Cooperative Extension Service and the library have worked together on an informational video project producing tapes about gardening, pet care, nutrition, and desert ecology.

California

The San Francisco Public Library has transformed a branch library into a Popular Media Center where for the first time patrons, especially those with poor vision, the blind, the hearing-impaired, and the physically handicapped, can view films or videotapes and listen to talking books or audiocassettes. The center has also established a video and sound production studio to produce local-interest audiotapes for the blind and signed videotapes for the deaf. For example, many elderly patrons with poor vision want the library to provide local-interest materials and materials in languages other than English.

Connecticut

The Danbury Public Library started video production in mid-1974. The library directs and cablecasts all access programs. One award-winning program, "Wonder Clock and Story Time," a reading readiness and appreciation program for children, is cablecast three times weekly from its own studio.

Georgia

The Tri-County Regional Library System in Rome, Georgia, produces and cablecasts programs directly from the library every day, plus three additional evening hours during the week. For example, every day at noon, a thirty-minute reference show called "Find Out" demonstrates search strategy, introduces standard reference materials, and cites the type of books that contain answers.

Illinois

The River Bend Regional Library System is producing twenty shows for cable television over a period of two years. The series of half-hour television shows is entitled "Other Things." Topics for the first year's programs are: a children's magic show, "The Delta Queen Riverboat," a puppet show and a program on making puppets, canning and freezing home-grown foods, helpful hints on housing, a children's program on pet care, non-traditional higher education, and winter outdoor activities.

New York

The Tompkins County Public Library in Ithaca produces a weekly Community Information Service. Short announcements about programs and services of nonprofit organizations are taped in the library and cablecast following the network newscast. Since January, over fifty community organizations have been publicized. In addition to programming for cable, the library tapes book and record reviews for regular showings on WSKG-Public Television on Mondays and Fridays during the 7 to 10 p.m. "Newscenter" telecast. The library also supplies WSKG with feature tapes on major local events.

Since December 1976, the Mid York Library System has programmed videocassettes over the local cable station. Starting with thirty-five half-hour programs purchased from the Public Television Library, Mid York has added programming from local sources and presently broadcasts one hour a week. The recent purchase of a color video camera should expand present cable air time and taping even more for public use.

New Jersey

The Ocean County Public Library's half-hour weekly cable programs are reaching fifteen communities in New Jersey. For instance, *Community Awareness* is the topic for the fourth week. Each month, fifteen minutes of this program are devoted to interviews with club and organization members who explain their programs, activities, and events they sponsor. In the rest of the library show, local government officials are interviewed.

Pennsylvania

Since May 1975, the Altoona Public Library has been producing weekly programs for showing over cable television. Topics range from historical and community service subjects to highlighting specific organizations. Most programs are thirty minutes in length and are shown three times a week. Videocassettes produced range from parenting to a documentary on making banjos, and from ETV programs taped off-the-air (with permission) to a historical program

on railroad steam engines. Citizens and groups can also make their own programs.

Wyoming

The Natrona County Public Library, in the city of Casper, serves an area of approximately 300 square miles. The library has been using videotape programming and producing its own tapes since the early 1970s. Natrona has its own in-house production studio and a two-channel cable system. One channel is strictly for programming; the second acts as a video reference. The channel used for library programming falls into two categories: public information and library information. To use the video reference service, the patron simply phones in a request for visual information and then tunes the television set to the channel for a reply. The service operates seventy hours per week.

American Library Association
Washington Office
November 1977

THE IMPACT OF COMMUNICATIONS TECHNOLOGY ON THE RIGHT TO PRIVACY

Public, research, and special (e.g., medical) libraries in the U.S. are today increasingly involved in the development and utilization of electronic data bases and communications media. Current trends make it evident that future patterns of library service to businesses, research groups, organizations, and individuals will utilize electronic communications media in ways already common (e.g., bibliographic searches, accessing research reports, and abstracts) and in ways still to be developed.

Privacy concerns in relation to libraries are twofold. In practical terms, library patrons in, for example, extremely competitive research areas will want to be assured that the topics of their investigations will remain confidential. In legal terms, library patrons have—in the view of the American Library Association—a constitutional right to privacy in their use of library resources.

The First Amendment protects the right to receive information, *Virginia State Board of Pharmacy* v. *Virginia Citizens Consumer Council, Inc.*, 425 U.S. 748; *Martin* v. *City of Struthers*, 319 U.S. 141, 143 (1943). It protects the anonymity of the author, *Talley* v. *California*, 362 U.S. 60 (1960); the anonymity of members of organizations, *Gibson* v. *Florida Legislative Investigation Committee*, 372 U.S. 539 (1963); *Bates* v. *City of Little Rock*, 361 U.S. 516 (1960); *NAACP* v. *Alabama*, 357 U.S. 449 (1958); the right to ask persons to join a labor organization without registering to do so, *Thomas* v. *Collins*, 323 U.S. 516 (1945); the right to dispense and to receive birth control information in private, *Griswold* v. *Connecticut*, 381 U.S. 479 (1965); the right to have controversial mail delivered without written request, *Lamont* v. *Postmaster General*, 381 U.S. 301 (1965); the right to go to a meeting without being questioned as to whether you attended or what you said, *DeGregory* v. *Attorney General of New Hampshire*, 383 U.S. 825 (1966); the right to give a lecture without being compelled to tell the government what you said, *Sweezy* v. *New Hampshire*, 354 U.S. 234 (1957); and the right to view a pornographic film in the privacy of your own home without governmental intrusion, *Stanley* v. *Georgia*, 394 U.S. 557 (1969). In light of these authorities, the American Library Association believes that the First Amendment protects the privacy of library patrons in their utilization of library resources, at least in the absence of a showing of a clear and present danger which threatens an overriding and compelling interest.

Due to the growing concern of American citizens regarding privacy, the American Library Association has by formal policy urged that libraries safeguard access to all records regarding patrons' uses of library resources (see attachment to this paper). In an era of data automation, interconnected data bases, etc., this concern for privacy will require, *inter alia*, the obliteration of patron-identified records of uses of library resources when those records are no longer needed for bona fide library purposes; careful selection of the means of patron identification (e.g., other than the social security number); and safeguards to eliminate unwarranted intrusion on and monitoring of communications channels used in library research.

For the above reasons, the American Library Association strongly supports the recommendations contained on p.614–15

of the option paper. Federal guidelines established by law will help assure uniform practices of privacy throughout interstate communications networks involving libraries.

ATTACHMENT

Policy on Confidentiality of Library Records

The Council of the American Library Association strongly recommends that the responsible officers of each library in the United States:

1. Formally adopt a policy which specifically recognizes its circulation records and other records identifying the names of library users to be confidential in nature.
2. Advise all librarians and library employees that such records shall not be made available to any agency of state, federal, or local government except pursuant to such process, order, or subpoena as may be authorized under the authority of, and pursuant to, federal, state, or local law relating to civil, criminal, or administrative discovery procedures or legislative investigative power.
3. Resist the issuance or enforcement of any such process, order, or subpoena until such time as a proper showing of good cause has been made in a court of competent jurisdiction.†

†Note: Point 3, above, means that upon receipt of such process, order, or subpoena, the library's officers will consult with their legal counsel to determine if such process, order, or subpoena is in proper form and if there is a showing of good cause for its issuance; if the process, order, or subpoena is not in proper form or if good cause has not been shown, they will insist that such defects be cured.

Adopted January 20, 1971; revised July 4, 1975, by the ALA Council

CONCLUSION

To live a productive life in today's world, information is not a luxury but a necessity, and all indications are that information will be even more of a necessity in the world of tomorrow. The amount of time, effort, and money citizens must expend to gain access to all types of information depends to a large extent on the communications policy of the United States. Because the strength of the nation is based upon an informed citizenry, we believe equal access to information is a matter of public policy. We urge that you consider this a major goal as you update the Communications Act of 1934.

We wish to acknowledge the contributions of the following members of the American Library Association's Information Science and Automation Division and of ALA staff in the preparation of this statement:

Public Broadcasting: Kenneth E. Dowlin, director, Pikes Peak Regional Library District, Colorado Springs; Harold E. Wigren, telecommunications consultant, Washington, D.C.

Domestic Communications Common Carrier Policy: Sara Case, legislative associate, ALA Washington Office; Ronald F. Miller, executive director, California Library Authority for Systems and Services; Mary Jane Reed, associate director for research and planning, Washington State Library.

International Telecommunications: Joseph Becker, president, Becker and Hayes, Inc., Los Angeles; Ken Winslow, manager, Video Program Service, Public Television Library.

Options for Cable Television Regulation: Brigitte L. Kenney, associate professor, Graduate School of Library Science, Drexel University.

The Impact of Communications Technology on the Right to Privacy: Judith F. Krug, director, ALA Office for Intellectual Freedom.

We appreciate this opportunity to present to the subcommittee the views of the American Library Association on communications policy.

Technical Communications

Innovative Uses of OCLC Records

Several months ago, OCLC announced to its participating libraries that they would produce magnetic tapes of an institution's catalog records and that these machine-readable records would be in the Library of Congress MARC II communication format. The tapes would be produced either weekly, biweekly, monthly, quarterly, or semiannually, and for a short time all retrospective records would be available.

The Caltech library staff felt that this opportunity should not be lost. Some staff felt that the machine-readable tapes were redundant, since the records were available on-line and in printed form in the card catalog. However, the prevailing view maintained that machine-readable records could be easily manipulated to provide a variety of output products and that this was a unique opportunity to develop new products for our users.

The first product was a humanities/social sciences acquisitions list. Users currently had to come to the library each week to learn about new books. OCLC tapes on a monthly or quarterly basis could be used to generate printouts sub-arranged by library, call number, author/editor, or subject. These lists could then be sent to users. The second product was a keyword-out-of-context (KWOC) index to the titles of all new cataloged books. Libraries have long suffered from archaic and inappropriate subject headings, and a KWOC title index would offer users a more up-to-date access to our holdings. A further benefit would be the ability to check new orders by title. The acquisitions system at Caltech requires that books be ordered by OCLC-main entry. Preliminary checking of the KWOC index would quickly identify duplicates.

In developing the computer programs necessary to translate the OCLC tape record, it was felt that brevity and simplicity were paramount. We decided that the branch library, call number, main entry (or added entry for personal name, corporate name, conference, or uniform title), and title would be sufficient.

The initial output is shown in example 1.

Two *T* cards and one *A* card are required in our KWOC program (Bindex—which was written at Caltech). The *A* card identifies keywords. The *T* card is text (up to a maximum of 132 characters). A sample entry from the acquisitions list is shown in example 2 and from the KWOC Index in example 3.—*Dana L. Roth, Sciences Librarian, Millikan Library, California Institute of Technology, Pasadena.*

```
T  HUM  HD8776  .D49  SANDBROOK, RICHARD, ED.                    000
T  THE DEVELOPMENT OF AN AFRICAN WORKING CLASS                  001
A  THE DEVELOPMENT OF AN AFRICAN WORKING CLASS
```

Example 1. Sample Entry Derived from OCLC Tape.

```
Humanities
HUM  HD8776  .D49  SANDBROOK, RICHARD, ED.
THE DEVELOPMENT OF AN AFRICAN WORKING CLASS
```

Example 2. Sample Entry from the Accessions List.

```
African
HUM  HD8776  .D49  SANDBROOK, RICHARD, ED. THE DEVELOPMENT
OF AN AFRICAN . . .
```

Example 3. Sample Entry from KWOC Title Index.

Network Development Office

National networking activities continued to be a dominant factor in the work of the Network Development Office. The recently reorganized Library of Congress Advisory Committee held its first meeting in Washington on November 28–29. The activities of its predecessor, the Network Advisory Group, have been recorded in earlier reports. The committee, which is composed of senior representatives of library networks and systems, was established by the Librarian of Congress to advise him on various issues concerning the Library's role in the evolving national library and information service network proposed by the National Commission on Libraries and Information Science in its program document. This first meeting resulted in positive steps to create the means by which critical issues can be identified and articulated for subsequent action by the committee or the Network Development Office, which is acting as the interim coordinating agency for the national library bibliographic network at the recommendation of the committee.

One of the most important issues facing the committee is the determination of a legal and organizational structure for the national library bibliographic network. A subcommittee had been formed by the predecessor of the committee, the Network Advisory Group, and at the subcommittee's second meeting in October, there was considerable discussion of a draft work statement prepared to solicit contractual support for this effort. After deciding that the work statement needed further refinement, the subcommittee agreed to meet again early in 1978 to review the revised work statement. If the latter is satisfactory, it will be submitted to the parent Network Advisory Committee for its approval. At its November meeting, the Network Advisory Committee also agreed to allow nonexclusive review of the work statement by the Council for Computerized Library Networks, whose members represent a broad segment of the library community. The director of the Network Development Office is a member of the subcommittee on legal and organizational structure.

The advisory committee's predecessor also established a Network Technical Architecture Group, composed of technical computer representatives from several library network organizations, to design a technical network architecture. The group has outlined a series of "implementable" projects to provide concrete data that could be incorporated in the design of the network architecture. One such project was to extend the services currently provided to the Research Libraries Group by the Library of Congress to other systems. (See *LC Information Bulletin* for October 21, 1977, p.718–19). In extending these services to other systems, the Library of Congress and the Network Technical Architecture Group, together with other library systems, have submitted a joint proposal to a funding agency to obtain support for areas that cannot be covered by the individual institutions.

In the meantime, the Network Technical Architecture Group has turned its attention to more specific aspects of the design of the network architecture. Several areas of concern have been identified, including technical design, data base configuration, administration, implementation strategy, and operational network considerations, which must be investigated in terms of their effect on the architecture. Technical alternatives are being specified, together with network functional requirements, performance criteria, and a set of tasks to lead toward a pilot test network. Meetings of the Network Technical Architecture Group, the Network Advisory Committee, and its subcommittee on legal and organizational structure have been funded by the Council on Library Resources.

Concurrent with the activities of the Network Technical Architecture Group, the National Commission on Libraries and Information Science (NCLIS) in conjunction with the National Bureau of Standards (NBS) had established a Task Force on Computer Network Protocols. The task force's final report was published in December and specifies an applications level

protocol for use in the national library and information service network. This protocol allows applications on one computer to send and receive messages from an application at a different computer. The protocol is important in the work of the Network Technical Architecture Group, and several projects to implement, refine, and use the protocol to interconnect library systems have been identified by that group.

The final report of the NCLIS/NBS task force had noted that the protocol specification is only the first step in creating an "implementable" protocol and that continued work was necessary in a test environment to evaluate or refine the protocol with respect to lower level protocols, site-specific problems, exception handling and error recovery procedures, security and protection of systems and data, and other technical areas. The Library of Congress through the Network Development Office is continuing its support of this protocol effort by funding a project with the New York Public Library to develop an implementation plan for testing and refinement of the protocol.

Other projects related to the evolving national network include ones on the role of authority files in the national network and the configuration of the national data bases. These two projects are proceeding in tandem at this time. The authority file study was actually begun in the fall of 1976 under contract to the National Library of Canada with funds provided by the National Commission on Libraries and Information Science. Although this first phase was begun with an emphasis on the role of authority control in a network environment, it became apparent very quickly that the problem of authority control could not be separated from the problem of designing a network data base configuration. The final report of this first phase, which was written by Edwin Buchinski, reflects this change in emphasis with the title "Initial Considerations for a National Network Data Base." This report is being prepared for publication by the Network Development Office.

This report has also been the basis for another proposal to funding agencies to continue the work leading to the design of the network data base configuration. NCLIS has agreed to fund a portion of this continuation effort. Because the total configuration must include location records as well as authority and bibliographic records, the problem of location data must also be considered. The final report of a project addressing this issue is being prepared for publication by the Network Development Office under the title "A National Location Data Base and Service." One of the points made in this study, which was performed by Butler Associates with funding from the Council on Library Resources, is that the requirements for consistent, authoritative cataloging data in a network union catalog should not inhibit the growth of a location data base whose requirements are far less stringent.

As work progressed on all of the projects mentioned above, it became obvious that communication was hindered by the lack of a common vocabulary. In June, the Network Development Office began a project to compile a glossary of library networking terms and contractual support from Dataflow Systems. A draft of the glossary has been completed and is being examined by representatives of various networking groups. The glossary is scheduled for completion early in 1978.—*From* LC Information Bulletin *January 13, 1978.*

News and Announcements

University of Pittsburgh Professor Completes Study of Self-Planned Learning in America

Four out of every five American adults plan and carry out one or more independent learning projects each year according to a national study *Self-Planned Learning in America*, recently completed at the University of Pittsburgh. First of its kind to be conducted on a national scale, the study examined the learning behavior of a representative sample of the total adult population of the United States. Patrick R. Penland, professor in Pittsburgh's Graduate School of Library and Information Sciences, directed the study sponsored by the United States Office of Education. He believes that the findings, which reveal that some American adults engage in as many as eighteen independent learning projects in a single year, have important implications for both formal and informal educational institutions and agencies.

Among the major findings of the year-long national survey of independent adult learning in America are:

- Seventy-five percent of adult learning projects are self-initiated.
- Self-initiated adult learners can be found at all social, economic, educational, and occupational levels in the population. Traditional demographic characteristics, therefore, are not always useful in predicting those individuals most likely to engage in systematic learning projects.
- Self-initiating adult learners are highly "goal-oriented," and patterns of self-directed learning are very individualistic. Adults attempt to draw on a wide range of community resources. The study suggests the need for interdisciplinary community learning resource teams—consultants, librarians, and other information professionals—to provide "shopping center" access to the delivery of human educational services.
- Adult learners often feel a strong need to establish the pace and control the character of their learning experiences. For this reason, only one out of five adult Americans enrolls in formal school or college courses where responsibility for guiding one's own learning must be relinquished to a teacher or group leader.
- The desire of individuals to set their own learning pace and to explore their own styles of learning, rather than submit to formal course-oriented experiences, was cited far more frequently as a reason for avoiding formal adult education courses in schools and colleges than were such factors as lack of time, transportation problems, or cost, considerations commonly believed to be significant deterrents to enrollment.
- Individuals develop their own learning patterns, sometimes obtaining help from human learning consultants, group planners, and such nonhuman resources as television courses, programmed instruction, or workbooks, but often by trial and error.
- The time devoted to a single learning project can range from 1 to more than 900 hours. The average time for a single adult learning project was 156 hours, which is more than three times longer than the number of contact hours in a typical semester-length college course.
- Adult independent learners most often prefer to study at home. On-the-job training also ranks high. Formal classroom settings were the least acceptable study location among adults. The study predicts that there is likely to be growing public sentiment for formal recognition and the award of "credits" for forms of study other than school and college courses.
- Nearly 60 percent of Americans have never made use of a library to pursue a self-initiated learning project. Only 14 percent of the adults surveyed reported that they used any library on a regular basis. An earlier study directed by Pen-

land, *Librarian-Client Learning Projects*, revealed that even this use of the library is frequently limited to simple assistance in locating findings or getting directions. Libraries are not generally perceived as a significant source of help by a majority of individual adult learners.

- The study establishes the concept of the "learning episode" as the basic unit around which the development of a learning project is constructed. Learning episodes may be as short as ten minutes or as long as an hour. Among respondents who perceived themselves as continuing learners, 73 percent had deliberately looked up some information within the previous seven days. The amount of time spent in this process varied but in 87 percent of cases was limited to one hour or less.
- Self-motivated adult learners differ in a number of significant ways from those who do not initiate learning projects or whose learning experiences are limited to formal courses or similar activities. For example, the frequency of memberships in organizations is significantly higher for adult continuing learners than for the adult population as a whole. By contrast, over half the nonlearners belonged to no organization whatsoever.

Self-Planned Learning in America explores in detail the behavioral aspects of learning by the average citizen, to provide an empirical basis for the development of the role of the librarian as a community learning consultant. Respondents were selected from the total population of the United States, eighteen years of age and older, in November 1976 by a modified probability sample designed by Marketmath, Inc., for the Opinion Research Corporation of Princeton, New Jersey, which conducted the pilot testing and the survey data collection. Data handling was completed by the University of Pittsburgh's Center for Urban Research. As principal investigator, Penland was assisted by a team of social psychologists, a national advisory committee, and consultants from the Communications Media Research Center at the Graduate School of Library and Information Sciences of the University of Pittsburgh. Support for the study was provided through a grant from the Library Research and Demonstration Branch, Office of Libraries and Learning Resources, U.S. Office of Education under Title II B of the Higher Education Act.

Copies of the final project report, *Self-Planned Learning in America*, are now available at $6.50 each from the University of Pittsburgh Book Center, 4000 Fifth Avenue, Pittsburgh, PA 15213.

New Computer for Lockheed's DIALOG Service Increases Capacity, Makes Searching Easier, Quicker

A new computer for Lockheed's DIALOG information retrieval service has increased the capacity of the system and has made retrieval of information easier and faster.

The computer—an IBM 370/165—has nearly four times the processing speed of the previous configuration, reduces response time, and allows for a greater number of connections.

The DIALOG service formerly used two IBM 360/65 computers, one primarily for TYMNET telecommunications service users and one for those using the TeleNet network. Since all access ports now connect to one computer, a user disconnected by TYMNET may redial through TeleNet and vice versa.

Another advantage of the new configuration is that one of the 360/65 computers from the former arrangement has been retained and is dedicated to file updating and processing, accelerating those functions as well as providing an emergency backup for the new computer.

On-Line Catalog Conversion Commercially Available

The Tacoma Public Library is currently converting its card catalog to machine-readable form by means of a unique minicomputer approach. Tacoma's problems were the same as those facing many libraries today: how to make the change from the manual catalog to one that can be maintained via computer, and how to provide the bibliographic information necessary for the new automated circulation system. They feel the approach developed for them provides a considerable savings both

in the cost of developing the record, as well as the library's personnel time.

DataPhase Systems, Inc., of Kansas City, Missouri, undertook the joint development of the method for providing the access to a large bibliographic file on the minicomputer, which DataPhase is providing as part of their automated library information systems for Tacoma Public Library, operational in April 1978.

Tacoma Public selected Blackwell North America through an open bid process as the supplier of the 1.4 million-title file of MARC records. Blackwell also provided technical advise and support for the project.

With a computer disk storage of almost 600,000,000 characters on the Tacoma computer configuration, DataPhase was able to load much of the MARC records for matching purposes. The task of conversion is being carried out by people funded through the Comprehensive Employment and Training Act, using twenty computer terminals on-line to the minicomputer.

The terminal operators match records by information available from the book: Library of Congress card number, International Standard Book Number, or the title and main entry author information. A match on any of the above allows the operators to accept the record entry and add local copy level data, such as call numbers, price, and location. The automatic entering of an optically readable item number (which is also people readable) completes the conversion.

Blackwell North America has statistically sampled the file and determined that approximately 80 percent of Tacoma Public's records should be already cataloged on the data base. For the remaining titles, DataPhase is providing an on-line interactive (question and answer) cataloging module for the entering of what can be full MARC records.

Tacoma estimates time savings of more than 50 percent of the staff's time over traditional methods of conversion while only having to handle most records once. Both Blackwell and DataPhase have subsequently decided to make this approach commercially available to other libraries for use in conjunction with or as an alternative to existing conversion systems.

Sources from both companies suggest the cost savings for full MARC records can be substantial, while giving the library complete control over their own project.

Blackwell North America is located at 10300 S. W. Allen Blvd., Beaverton, OR 97005. DataPhase Systems, Inc., is located at 4528 Belleview, Kansas City, MO 64111.

Full Range of Conversion Services Available from CLSI

CLSI has announced the availability of a variety of conversion services that facilitate the transition to on-line circulation. Conversion is the process by which existing patron, title, and item data are entered into a fully on-line system. The objective of the conversion strategy is to enter a sufficient amount of data prior to the start-up of on-line operations to ensure efficient system performance. Several considerations determine the method of conversion selected by a particular library—its technical, financial, and personnel resources and the method of data collection being replaced.

CLSI now offers three conversion services that eliminate the need to enter bibliographic information manually at the keyboard. These conversion services, which are closely monitored to ensure that high quality bibliographic information is delivered, include:

- Archival tapes from a bibliographic reference service to CLSI format—CLSI will arrange information on the library's archival tapes to conform to the CLSI format.
- Custom conversion—CLSI will convert a library's shelflist to the CLSI record format on tape, including microfilming the shelflist, keying and verifying the shelflist, and creating the input tape for the LIBS 100.
- Preloaded files—Working with the library staff, CLSI representatives will establish a library profile designating the strong subject areas of the collection and years of publications represented. A data base will be created that will match the library's profile.

In addition to these new conversion

services, CLSI has developed several other methods of converting bibliographic information and patron registration files to the LIBS 100 System.

- The LIBS 100 Circulation/Conversion Station—Libraries utilizing a variety of alternate data input media, such as eighty-column punched cards used in batch-oriented data processing systems and Hollerith badges often used for patron identification in academic libraries, can transfer this machine-readable data to the LIBS 100 data base via the LIBS 100 Circulation/Conversion Station. This method enables libraries to take advantage of other automation efforts that have already produced machine-readable data.
- Data loading via LIBS 100 Magnetic Tape Facility—A library that has patron and/or title data on tape or cards in a format suitable for conversion into CLSI's record format may utilize CLSI's Magnetic Tape Facility to load the data onto disc. In addition, after the patron and/or title files exist on disc, the library may use the Magnetic Tape Facility to update these files on a regular basis.
- Master title file—CLSI will arrange for a library to obtain a master data base from a similar LIBS 100 library that contains the most popular titles in public libraries. The assumption underlying this technique is that the library will have a high percentage of titles available from the LIBS 100 library's master data. The title purge function of the LIBS 100 enables the library to purge the titles on the master data base that are not contained in the library's collection. This method eliminates the manual keying of title information.
- Conversion "on-the-fly"—The library enters a minimum amount of patron and item information at checkout. Complete patron and item information is entered at a later time. This method, utilizing the CLSI two-part bar-encoded labels, permits patrons and books to be processed rapidly by the system at checkout—whether or not they have previously utilized the library or circulated on the automated system. Rapid and efficient keying is facilitated by special interactive dialogues, which have been developed based upon the actual use of the LIBS 100 in more than 100 public, academic, and special libraries across the United States and Canada.

CLSI is engaged in a continuous program to improve conversion techniques and eliminate the need for manual conversion. For additional information contact: Director of Communications, CLSI, 81 Norwood Avenue, Newtonville, MA 02160.

Complete Automated Library Systems from DECICOM

Decicom Systems, Inc., has announced a full product line for automating library circulation control. The major functions of the system are (1) in-and-out book checking, (2) reserve books, (3) renewal, (4) overdue notices and delinquent borrowers, (5) inquiries (by author, title, subject), (6) file maintenance, (7) interlibrary loan. Systems are available to automate all of these functions or a selected few, depending on a particular library's needs.

Among major options offered in the Decicom library system are "time-shared" and "stand-alone." On a time-sharing basis, a library can opt for batch or on-line hookup with the central library system. Batch transactions are recorded at the library on storage media (diskette or magnetic tape) throughout the day. These transactions are sent periodically to the time-share computer, by mail or through communications lines. At the time-sharing center, the storage media is processed and reports sent back to the library through communications lines. This process can be done weekly, daily, or as the library demands.

An on-line/time-shared combination enables the time-shared option to operate as a stand-alone system. The computer resources are shared between libraries, usually at the county level, and information is more readily available whenever it is needed.

With the stand-alone option, all circulation functions are available. The computer is located in the library, allowing its information to be accessible at any time. The computer is programmed in English-type language, allowing the library to add its own applications, as a need may arise.

(These may include payroll and attendance functions.)

Prices for the Decicom automated library systems vary from $25,000 (for "batch") to $100,000 (for "stand-alone"). The systems are so designed that a library can start out with the least expensive "batch" system and gradually add to it as need and budget allow.

The Decicom automated library system employing the time-shared option has already been installed in several libraries and is operating successfully.

For more information on Decicom's full product line for automating library circulation control, contact Decicom Systems, Inc., 250 Adams Boulevard, Farmingdale, NY 11735.

Dallas Junior College District Implements New Concept in Library Automation

A new approach to a total library information system has been implemented at the Dallas County Community College District, Dallas, Texas. The project is the pilot installation for the commercial offering of the Automated Library Information System by DataPhase Systems, Inc., of Kansas City, Missouri.

The system was conceived using the data processing techniques of data base management, which allows for all information, or files of data, to be on-line and available to all users at all times. Changes to the files are immediately updated and the status is current at all locations.

The structure of the bibliographic file allows for variable length of fields within variable-length records, a technique used to achieve MARC compatibility, and offers the option of maintaining the card catalog on the computer. The colleges will be updating their COM catalogs as a direct by-product of the automated system.

The district uses optical character recognition as the method of identifying materials with human-readable letters and numbers. This offers the opportunity for machine-readable scanning by means of a hand-held optical scanner provided by DataPhase Systems, Inc. The system was designed to operate on a mini-computer, which can flexibly grow as the requirements of the district evolve.

The college district includes seven separate colleges as well as a technical services center. Paul Dumont is the director of the center and has the primary responsibility of serving the needs of the resource center at the individual campuses.

DataPhase Systems, Inc., is a company specializing in library automation. Since the award of the contract in Dallas one year ago, the company has been successful in developing programs at a number of other academic and public libraries. The offices are located at 4528 Belleview, Kansas City, MO 64111.

NMA Publishes 1978 Buyer's Guide

The National Micrographics Association (NMA) has announced the publication of its 1978 *Buyer's Guide to Micrographic Equipment, Products and Services.*

The reformatted sixty-four-page directory, distributed free, serves as a concise introduction and a continuing reference to many valuable sources of currently available products and services for both present and potential micrographic customers.

The 1978 *Buyer's Guide* lists NMA trade members in two sections: products and services. This year's *Guide* also features an easy-to-use, index-style listing.

In the product section, names of companies that actually manufacture a specific item are in roman type, while names of companies that act as distributors and/or dealers *only* of the product are italicized.

To assist the user in locating a service company in a particular area, a geographical listing of service companies only is included in the services section.

An alphabetical listing of trade members is included at the back of the *Guide*, followed by the index to products and services listed in the publication.

The 1978 *Buyer's Guide* includes data on manufacturers, dealers/distributors, consultants, service companies, micropublishers, and trade publishers in the micrographic industry. More than 200 NMA trade member companies are categorized in 22 product or service sections as part of NMA's ongoing service for information specialists.

Free copies of the 1978 *Buyer's Guide to*

Micrographic Equipment, Products and Services may be obtained by writing to NMA Publications Sales, 8728 Colesville Rd., Silver Spring, MD 20910.

Super Key

Super Key, a new company, was formed for the exclusive purpose of providing conversions of libraries' bibliographic files from hard copy or microfilm to machine-readable formats for catalog and circulation data base records.

The catalog conversion service features keying bibliographic data base access fields, such as LC card numbers, ISB numbers, author, title, or algorithms thereof, for extracting records and full MARC format(s) conversion of nonmatched records. Call number and owning library(ies) are also keyed.

Super Key has a systems librarian to supervise every library conversion to insure the most accurate shelflist keying.

To date, the union main entry file for Wayne-Oakland County Federated Library System and the shelflists for ten Nebraska community colleges have been keyed and matched to Library Interface Systems, Inc./General Research Center data base. Full record conversion of nonmatches is in progress for both files. For further information, write to Super Key, 1421 E. Wayzata Blvd., Suite 51, Wayzata, MN 55391.

Future of the Cornell University Catalog

Report number one of the Task Force on the Future of the Cornell University Libraries Card Catalogs has been published. The report culminates two years of study and discussion and is intended to initiate a dialogue with the entire libraries staff concerning the possible closing of the Cornell University Libraries' card catalogs.

The report recommends that catalogs be closed at the time the new cataloging code (*AACR* II) is adopted; that a new on-line system with a COM backup system be implemented; and that the automated system also manage the libraries' other processing tasks, i.e., acquisitions, serials, circulation/reserve, etc. The committee believes this report represents one of the most comprehensive documents so far published on closing the card catalogs in a major research library.

Copies of the report are available for $2 each, prepaid, from the Budget & Accounting Office, 234 Olin Library, Cornell University Libraries, Ithaca, NY 14853.

BALLOTS and RLG to Cooperate

Stanford University and the Research Libraries Group, Inc. (RLG), have jointly announced a decision by the RLG board of directors to use Stanford's computer-based bibliographic processing system, BALLOTS. According to the board's decision, all members of RLG will be using the system by the end of 1981, though some members may begin using BALLOTS by the end of 1978.

BALLOTS ("Bibliographic Automation of Large Library Operations Using a Time-sharing System") is an on-line library system originally developed by Stanford University for its own libraries and later expanded for use by other institutions. It provides bibliographic information storage and retrieval with file access through a variety of indexes designed for use by libraries and library systems of all types. BALLOTS is presently used by academic, public, research, and special libraries.

Stanford University has committed BALLOTS to develop a bibliographic network for research libraries at the national level and to develop a bibliographic utility for all libraries throughout the state of California.

New BALLOTS record, index, and file structures, being developed for use by the end of 1978, will expand and improve its services by making it even easier for users to locate desired materials in a growing data base of larger scope and by making plain what standards were followed in any cataloging data retrieved. Planned BALLOTS development also includes an improved system of computer-remote terminal communication, an acquisitions system to support a network of users sharing acquisitions data, cataloging authority control, and expanded support of card, book, and microform catalogs.

The Research Libraries Group was formed in 1974 as a cooperative activity of

large research libraries, including Columbia University, Harvard University, Yale University, and the New York Public Library. It is a private, nonprofit organization governed by a board of directors appointed by its members. Since its inception, RLG has conducted a number of planning studies and operational activities directed at reducing costs and improving services to library users through cooperation.

In 1974, RLG made a decision to use a computer system in order to establish and maintain a union catalog of machine-readable bibliographic information, whose uniformity would be controlled by common acceptance of cataloging practices and standards based on those used by the Library of Congress. This bibliographic uniformity would permit the RLG institutions to contribute unique cataloging records to an emerging national bibliographic data base and to support their cataloging and acquisitions work. RLG also desired a system that would permit library patrons to identify materials related to an author or subject and to determine in which library the materials could be found.

In its search for an appropriate computer system, RLG conducted a number of site visits to candidate systems and solicited responses to a request for a proposal. In addition, an independent consulting organization, BGS Systems, Inc., was engaged to conduct an analysis of functional and operational characteristics of the leading candidate systems.

Both parties view RLG's participation in the development and use of Stanford's BALLOTS system as an initial step in addressing the needs of research libraries and their users at the national level.

As both Stanford and RLG require a number of features still to be implemented in the system, RLG is committed to working with Stanford/BALLOTS in establishing development priorities and obtaining funding to support these efforts. The time required for augmentation of the system is the reason for setting December 1981 as the target date for full participation by all RLG members.

Cooperation between Stanford University and the Research Libraries Group can maximize earlier funding provided to both organizations for development work. Development grants have come from the Andrew W. Mellon Foundation, the Alfred P. Sloan Foundation, the National Endowment for the Humanities, and the Council on Library Resources.

Officers of Stanford and RLG are now drafting a memorandum of understanding to serve as the initial step in defining developmental and operational issues and the responsibilities of both parties. Both anticipate that other research libraries will adopt BALLOTS in the near future.

Public Library Surveys
CLSI'S LIBS 100 Users

A public library, in considering the purchase of a computerized system for automated circulation control, surveyed C L System's (CLSI) LIBS 100 users. Thirty-five libraries responded to the questionnaire, which:

- asked about any cost savings resulting from installation of the LIBS 100 System;
- questioned the performance of CLSI in the delivery of the software package, the performance of the equipment, the response to service needs, and the educational effort for staff involved with the system;
- recommended the minimum annual circulation necessary to justify automation;
- determined why the libraries decided to automate;
- evaluated other automated circulation systems available.

On the basis of experience, thirteen (86.7 percent) libraries replied that they experienced cost savings with the LIBS 100 and two (13.3 percent) libraries said they have definitely not experienced cost savings. Twenty libraries either did not respond to this question, as they were not yet on-line, or had other responses.

James Walton, circulation services librarian at Johns Hopkins University Milton S. Eisenhower Library, Baltimore, Maryland, stated: "We do not expect substantial savings this year (the first year after installation). But, over and above financial savings, we feel we are offering better service to more people. It is difficult to affix a cash value to that."

Table 1. Performance of CLSI

	Excellent	*Good*	*Problems*
Delivery of Software	46.7% (14)	50.0% (15)	3.3% (1)
Performance of Equipment	35.5% (11)	64.5% (20)	0% (0)
Response to Service	72.4% (21)	27.6% (8)	0% (0)
Educational Effort for Staff Involved with LIBS 100	41.4% (21)	48.3% (14)	10.3% (3)

Lewis & Clark Library, Helena, Montana, responded that the library had definitely experienced cost savings: "When we moved to a new facility in August 1976, we did not have to add personnel even though our circulation has increased 80%. We have also transferred some personnel (circulation workers primarily) to different parts of the operation."

On the basis of experience, the libraries qualified the performance of CLSI, as shown in table 1.

All thirty-five libraries responded that the major factor to be considered when contemplating automation is the inefficiency of the library's present circulation system—not the library's annual circulation. LIBS 100 Systems are operating in libraries with annual circulations ranging from as few as 20,000 materials per year to more than 1 million materials per year.

The general reasons for deciding to automate included: improving the libraries' methods for placing and trapping items on reserve; timely overdue and fine notices; improved circulation statistics; and locating library materials quickly. Nancy Hudson, CLSI project coordinator at the Clark County Library District, Las Vegas, Nevada, responded: "We are all very satisfied with the CLSI System—do not know how we ever operated without it. [The LIBS 100] completely automates the reserve book system, a service that had become so expensive with manual methods and so inefficient that we would have had to discontinue the service . . . . Overdue notices are produced *on time*, are accurate and legible . . . . Resource sharing among branches and agencies has been enormously improved . . . . Much better control has been established over delinquent patrons—before we had no control."

Most libraries agreed that none of the other automated services available offered the same range of services as CLSI. A representative from the Boise (Idaho) Public Library stated that although other circulation systems are available, "none offer the flexibility and the range of options that CLSI does." Frank Hannaway of the Providence Public Library responded: "We began with CLSI when they were not a household term. Since we have had CLSI, others have come into existence, but we would have still chosen CLSI."

For a detailed summary of the results of the survey contact: Mary Lou Cocci, C L Systems, Inc., 81 Norwood Avenue, Newtonville, MA 02160; (617) 965–6310.

VCCS News

Although a complete report of the meetings of the Video and Cable Communications Section (VCCS) of the Library and Information Technology Association, held at the ALA Midwinter Meeting in Chicago, will appear in a later issue of *CableLibraries*, here is a synopsis of the highlights.

Network Established

Based on the regions established by the NFLCP, regional coordinators for VCCS were selected in Chicago. They are the following:

Western Region: Roberto Esteves, San Francisco Public Library; Kandy Brandt, Seattle Public Library (Washington, Oregon, California, Nevada, Arizona, Alaska, Hawaii)

Mountain Region: (not selected yet) (Idaho, Montana, Wyoming, Utah, Colorado)

Southwest Region: Anne Hollingsworth, Texas State Library (New Mexico, Texas, Oklahoma, Louisiana, Arkansas)

Midwest Region: Marilyn Rehnberg, Rockford (Illinois) Public Library (North Dakota, South Dakota, Nebraska, Kansas,

Minnesota, Iowa, Missouri, Wisconsin, Illinois, Michigan, Indiana)

Southeast Region: Helen Karpinski, Memphis Public Library (Mississippi, Alabama, Georgia, South Carolina, North Carolina, Tennessee, Kentucky, Florida)

MidAtlantic Region: Lynne Bradley, Washington, D.C., Public Library (Virginia, West Virginia, Maryland, Delaware, New Jersey, Ohio)

Northeast Region: Linda Hillman, South Central Research Library Council, Ithaca, New York (New York, Vermont, New Hampshire, Massachusetts, Rhode Island, Maine, Connecticut)

The purpose of the network is multiple. Already it has been proposed that the network will be used to:

- Identify and build a complete file of names and addresses of all libraries involved in telecommunications work.
- Recruit members for VCCS and disseminate VCCS information.
- Report to *CableLibraries*.
- Take the pulse of librarians involved in telecommunication when issues arise.
- Supply information on telecommunications to interested people for the Governor's and White House Conferences.
- Facilitate tape exchange.
- Provide quick notice of grants and funds.

All libraries involved in telecommunications work are encouraged to contact their nearest regional coordinator so that they may become part of the VCCS network.

MEETINGS AND CONFERENCES

Pittsburgh Conference

The Graduate School of Library and Information Sciences, University of Pittsburgh, in cooperation with the National Commission on Libraries and Information Science, announces the 1978 Pittsburgh conference, "Toward the White House Conference: The Structure and Governance of Library Networks in Light of a Developing Technology." The conference will be held November 6–8.

The purpose of the 1978 conference is to examine library networks in terms of management and organization, standardization, criteria for network membership, choices among complex technological options, interrelationships among networks, measures of network performance, and financing.

The White House Conference on Libraries and Information Science, scheduled for October 1979, has as its theme "Equal Opportunity of Access to Information." The design and operation of library networks, to facilitate the sharing of both bibliographic data and library materials as well as human expertise, is central to the equalization of access to information for all citizens and thus a critical element in the formulation of a national information policy. The Pittsburgh conference will provide an opportunity to participate in the definition of problems that may constitute a significant part of the agenda of the White House conference.

Among the nationally recognized leaders participating in the conference will be: Henriette D. Avram, Library of Congress; Joseph Becker, Becker & Hayes, Inc.; John W. Bystrom, University of Hawaii; Melvin S. Day, National Library of Medicine; Ervin J. Gaines, Cleveland Public Library; Vincent E. Giuliano and Susan Crooks, A. D. Little, Inc.; Robert M. Hayes, University of California at Los Angeles; Dick Hays, U.S. Office of Education; James H. Kennedy, AMIGOS; Donald W. King, King Research, Inc.; Beverly P. Lynch, University of Illinois at Chicago Circle; John P. McDonald, University of Connecticut; Anthony W. Miele, Alabama State Library; James P. Riley, Federal Library Committee; Stephen R. Salmon, University of California; Edward E. Shaw, Stanford University; Charles H. Stevens, SOLINET; Roger K. Summit, Lockheed; Roderick G. Swartz, Washington State Library; and William J. Welsh, Library of Congress. Participants from the National Commission on Libraries and Information Science will include Frederick H. Burkhardt, Alphonse F. Trezza, and William Mathews. Participants from the University of Pittsburgh will include Thomas J. Galvin, Allen Kent, Patricia B. Pond, K. Leon Montgomery, James G.

William, Sara Fine, and Roger Flynn.

The speakers will share their knowledge, experience, and perceptions through presentations and discussions that will review network anatomy, objectives, technology, and governance, their purposes, expectations, and alternatives for the future.

Five position papers will be prepared before the conference is held and will be sent to all advance registrants before they arrive in Pittsburgh. The final program, with topics, speakers, and registration information, will be distributed in the near future. Early inquiries may be directed to: Allen Kent, Distinguished Service Professor, Graduate School of Library and Information Sciences, University of Pittsburgh, 801 L.I.S. Building, Pittsburgh, PA 15260.

Input

To the Editor:

With respect to the December 1977 editorial, certainly it is appropriate for the Reference and Adult Services Division to contain a Machine-Assisted Reference Section (MARS). As this year's projected total of 1,500–1,700 MARS members indicates, computer-based services have rapidly moved into the mainstream of reference librarianship, and RASD would have had to respond to that overwhelming fact no matter what ISAD had chosen to do. So there really is no reason for ISAD members to feel that they lost some kind of a contest, or that they have now forfeited all chance of a constructive role in this area of library work.

In fact, when the current reorganization of ISAD into LITA is settled, MARS would be most interested in exploring appropriate links to LITA in at least the following areas: hardware (especially terminals), telecommunications, and software (both search programs and housekeeping programs such as billing).

Reference librarians using on-line systems frequently have to deal with new options, products, services in these three fields. Since their mission is to be reference librarians, they frequently have neither the time nor the desire to become as expert therein as most LITA members are. We are anxious to cooperate with LITA and would welcome your suggestions as to the best mechanisms.

Sara D. Knapp, Chair
MARS Organizing Committee

To the Editor:

The March 1978 article by Pierce and Taylor comparing costs of OCLC/SOLINET, BALLOTS, and VPI manual cataloging does an irreparable disservice to OCLC and to SOLINET. It may also harm BALLOTS; I really don't know. The harm is in the inaccurate information, cited as coming from qualified sources but given wrongly.

Let me refer to the worst example I see. On page 16, the "Entrance Fee" for SOLINET is given as "$27,000." The memorandum cited in the article quite clearly says the fee is "the *greater* of $2,000 or 1½% of one year's library materials budget." SOLINET also provides its "Cost Estimate Calculator" and "Cost of SOLINET Membership" sheets to prospective members, which further clarify and confirm this point. All of these documents were available considerably before the December 14, 1977, revision.

VPI paid $19,000 to join SOLINET. *No member has paid as much as $27,000 to join!* Most of the libraries that join SOLINET pay the minimum fee—$2,000. I could defend this or any fee, but that is not my purpose. My purpose is to say that such an error in the premise is certain to make the conclusion deeply suspect.

I know that, but the innocent, trusting reader of *JOLA* thinks you and your referees have certified the article and he/she will trust that the cited fee is correct. Was the article refereed? Did anyone from OCLC or BALLOTS confirm the data? I know that it wasn't checked with me or with anyone here. Too bad. We were sent the February 1976 issue of the study in August of 1976. The given membership fee there—$22,500—is closer to the truth, but it also states that it was based on the book budget.

Can you imagine with what glee the purveyors of non-OCLC services will show this article to their candidates? Can you imagine what other OCLC networks will say about SOLINET fees as detracting from *their* activity since SOLINET is alleged to be typical? Small libraries will turn away in horror. Large libraries will suspect that SOLINET is wealthy and becoming more so at their expense. I can't fathom why Pierce and Taylor have per-

petrated this error and why they weren't caught in the editorial process.

Our own records of *actual* expense show costs for service of real members who are currently using the OCLC system (and applying the same factors as Pierce/Taylor) vary from $1.86 upwards. Perhaps in time we'll see an article on actual costs—one in which VPI looks at its costs as real users.

It's really too bad that *JOLA* has—in this article—done us such a disservice.

Charles H. Stevens
Executive Director
SOLINET

Articles submitted to JOLA *are subjected to a careful peer review process, which we believe to be adequate. Review of the article in question was especially thorough. Review, of course, is not censorship, nor can it extend to the basic data contained in an article. The following letter from the authors responds in more detail.*

WDM

Dear Mr. Stevens:

In reference to your letter to the editor, we hasten to explain what you perceive to be a discrepancy between the entrance fee paid by VPI and SU to join SOLINET and the entrance fee calculated in our article comparing projected costs of a potential implementation of both SOLINET/OCLC and BALLOTS to the actual costs of our manual monographic cataloging operation. The data we used as the basis for the calculations are taken from our 1976–77 fiscal year. During that year our library materials budget (which we call our "book budget") was $1,800,000. Had we actually joined SOLINET in 1976–77, we would have expected to pay $27,000 for an entrance fee. We chose to use this figure to maintain consistency in the projected costs by basing them all on the same time period. Had we chosen to use the actual entrance fee of $19,000 paid by VPI & SU when our library materials budget was $1,200,000, we would have been forced to explain that fortunate historical accident resulting in supplemental appropriations to join SOLINET, and we would have been open to the criticism of mixing apple-time with orange-time in our calculations.

We wish to stress again that these costs are estimated costs, based on 1976–77 data and the best estimates we could make. In fact, we felt uncomfortable enough with the estimations to calculate three possible values for both SOLINET/OCLC and BALLOTS.

We sincerely hope that the error you perceive in our premise has been explained and that, in any case, it, or any other minor inaccuracy, does not make our conclusions suspect; these conclusions form the foundation of our belief that automated cataloging systems *can* help control the rising costs of operating a library, especially personnel costs, if the application is carefully planned and executed to take advantage of the benefits of the system.

The goal of our paper was not to downgrade any one system but to provide a technique for a library, faced with the dilemma of choosing a cataloging alternative, to derive expected costs in a manner that would allow comparison. The method we propose for deriving expected costs can be used on any cataloging systems.

A. R. Pierce
Planning and Research Librarian
and
Joe K. Taylor
Assistant Planning and Research Librarian
Virginia Polytechnic Institute and State University

Book Reviews

Librarians and Online Services, by Pauline Atherton and Roger W. Christian. White Plains, N.Y.: Knowledge Industry Publications, 1977. 124p. $24.50. LC: 77-25275. ISBN: 0-914236-13-X.

This book addresses the questions of how and why librarians provide a relatively new reference tool, on-line reference services, to their patrons. The authors have limited their scope to on-line bibliographic reference services. No attempt is made to deal with on-line services involving the searching and manipulation of purely statistical data or machine-readable files of full-text documents. The point is well made that on-line reference services are ". . . a natural and important extension of the services the library makes available to its clientele." Atherton and Christian indicate that the changes that occur can be anticipated and managed intelligently for maximum benefit to the library, its staff, and its clientele. The way in which librarians react to these changes may determine the future successes or failures of libraries. The commitment that must be made to staff training, updating, and user education when implementing on-line reference services is stressed.

Nine chapters, (1) "Introduction and Overview," (2) "The Extent of Online Services in Libraries," (3) "Impact on Library Staff and Administration," (4) "Start-up Considerations," (5) "Financial Considerations," (6) "Modes of Operation and Service Procedures," (7) "Marketing and Promotion," (8) "Management and Control," and (9) "Conclusions and A Look At The Future," provide answers to questions often asked by those planning or considering implementation of on-line reference services.

There are tables throughout the book furnishing information on forms, checklists, and procedures used by various libraries already providing on-line services. The authors have made good use of previous studies and publications. The selected bibliography is a useful tool in itself. The questions of payment for library services (connect costs, etc., for on-line searching) and of marketing and promotion are addressed reasonably, without the hysteria sometimes associated with these topics. The differences in traditional library services and those possible with the aid of on-line services are clearly indicated.

The book is valuable for anyone launching into on-line services and especially good for those in academic or public library settings. The physical character of the book is not very pleasing, however. The typed format contains several spelling errors, typographical errors, and so forth. The tables are interspersed with the text. It would be preferable for them to be at the ends of chapters or in separate appendices.

Kay Rodgers
Library of Congress

The Measurement and Evaluation of Library Services, by F. W. Lancaster, with the assistance of M. J. Joncich. Washington, D.C.: Information Resources Press, 1977. 395p. $27.50. LC: 77-72081. ISBN 0-87815-017-X.

If you have been thinking about or have been assigned the task of evaluating part or all of your library's services, then F. W. Lancaster's latest book, *The Measurement and Evaluation of Library Services*, is "must" reading for you. If you have not been thinking about how to evaluate libraries, perhaps you should have been, and to help bring yourself up to date you should read Lancaster's excellent text as soon as possible.

"The book concentrates primarily on techniques that can be used to evaluate the public service of a library, preferably by means of reasonably objective procedures, although some consideration also is given to the evaluation of technical services" (p.ix). Lancaster does this by limit-

ing the scope of the book to ". . . how well the library satisfies the immediate tangible needs of its users. It deliberately excludes any consideration of the evaluation of libraries in terms of their broader, intangible, and largely unmeasureable 'benefits' to Society" (p.ix).

The volume "was developed primarily as a textbook . . . . It should be regarded largely as a survey and synthesis of, as well as a guide to, published literature in the field." It provides a comprehensive survey of the elements that form the quantitative school of library research. The first chapter ("The Evaluation of Library Services: An Introduction") provides definitions for many of the terms used by the writers of quantitative librarianship; a context in which to place the findings of their studies; a discussion of the problems public service institutions have with the meaning of their intangible objectives; and an explanation of the concepts of macroevaluation, microevaluation, and the measurement of exposure.

With the foundation laid by the first chapter, one can read any of the succeeding thirteen chapters, which appear in the following order: (2) "Studies of Catalog Use," (3) "Evaluation of Reference Service," (4) "Evaluation of Literature Searching and Information Retrieval," (5) "Evaluation of the Collection," (6) "Evaluation of Document Delivery Capabilities," (7) "The Range and Scope of Library Services," (8) "Evaluation of Technical Services," (9) "Evaluation of Automated Systems in Libraries," (10) "The Relevance of Standards to the Evaluation of Library Services," (11) "Library Surveys," (12) "Effect of Physical Accessibility and Ease of Use," (13) "Cost-Performance-Benefits Considerations," and (14) "Conclusion: Factors Affecting the Performance of Library Services."

Even if you are up to date on this vital area of library research, you will want to read the first chapter to see how Lancaster places this research into a context. The index provides a useful check to see if Lancaster has covered all of the researchers you know. The references at the end of each chapter are excellent and form a reading list for the beginner. The illustrations are of fine quality throughout the text.

In the final chapter, "Conclusion: Factors Affecting the Performance of Library Services," R. H. Orr's important article, "Measuring the Goodness of Library Services: A General Framework for Considering Quantitative Measures" (*Journal of Documentation* 29, no.3: 315–32 [1973]), is discussed along with the work of Rzasa and Baker ("Measures of Effectiveness for a University Library," *Journal of the American Society for Information Science* 23: 248–53 [1972]). The supporting tables in this chapter are very helpful. There are some ninety-four tables and sixty-five figures in the text.

At $27.50 a copy, not all librarians will purchase the book, but all should read and study this book, because it presents a great deal of useful research-based information about libraries and the ways in which people use them.

Neal K. Kaske
University of California, Berkeley

BOOKS RECEIVED

Collection Development Analysis Using OCLC Archival Tapes; Final Report, Glyn T. Evans, Project Director, State University of New York Central Administration, Office of Library Services, Albany, New York. Washington, D.C.: U.S. Dept. of Health, Education and Welfare, Office of Education, Office of Libraries and Learning Resources, 1977. 65p.

Book Theft and Library Security Systems, 1978–79, by Alice Harrison Bahr. White Plains, N.Y.: Knowledge Industry Publications, 1978. 128p. $24.50. LC: 77-25284. ISBN: 0-914236-14-8.

Information Systems and Networks: Design and Planning Guidelines of Informatics for Managers, Decision Makers and Systems Analysts, by K. Samuelson, H. Borko, G. X. Amey. Amsterdam: North-Holland, 1977. 148p. LC: 75-40169. ISBN: 0-7204-0407-X.

Video in Libraries: a Status Report, 1977–78, by Seth Goldstein. White Plains, N.Y.: Knowledge Industry Publications, 1977. 104p. $24.50. LC: 77-3554. ISBN: 0-914236-07-5.

BLACKWEL

A NAME THAT SPEAKS WITH AUTHOR

Right now!

To be more specific—fully automated Subject Authority Control according to LCSH/8.

By the time you read this advertisement, every COM and book catalogue produced for B/NA-customer-libraries will have been fully recatalogued from LCSH/7 to LCSH/8. And all the required cross-references will have been interfiled and verified.

All this will have been done at B/NA with B/NA software, the LCSH/8 database, and B/NA's experienced staff of librarians and editors.

The Subject Authority Control System will process computer cataloguing from such sources as: LC-MARC records, OCLC user records, CAN-MARC records, B/NA-MARC records, and other commercial vendor MARC-like records (which we first convert to MARC).

B/NA is the only vendor which can deliver this service to your library now.

Right now! Just as we deli
to LCSH/8) with LCSH
proven software. The sam
staff. The most up-to-date

B/NA technical services customers include college, research, state, public, and special libraries around the world.

To sum up, then, we would like to speak with you about some, or all of the following:

- shelflist conversion to MARC;
- database management of OCLC (or other network) records;
- book, (COM) fiche, or film catalogues;

and, of course,

- Subject Authority Control by LCSH/8.

Please call (503) 643-8423. Or write to Blackwell North America, 10300 S.W. Allen Blvd., Beaverton, Oregon 97005. Ask for Michael Moen, one of the twelve librarians on our technical services staff, all of whom are knowledgeable about MARC structure, COM fiche and film catalogues and, of course, fully automated Subject Authority Control by LCSH/8.