today's large academic libraries struggle, there is, nonetheless, room for criticism of library priorities.

This study must be viewed as only a first step (largely tentative and exploratory) in relating automation with service attitudes. It suggests that online systems may be associated with managers more positive in their view of the management role and more positive in their attitudes toward users than batch- and manual-system managers. Further research would be useful at this point to compare levels of automation (manual, batch, and online) with circulation-staff service attitudes or those of patrons using the systems.

REFERENCES

1. Laurence Miller, "Changing Patterns of Circulation Services in University Libraries" (Ph.D. dissertation, Florida State University, 1971), p.iii.
2. Ibid., p.149.
3. Robert Oram, "Circulation," in Allen Kent and Harold Lancour, eds., *Encyclopedia of Library and Information Science*, V.5 (New York: Marcel Dekker, 1971), p.1.
4. William H. Scholz, "Computer-Based Circulation Systems—A Current Review and Evaluation," *Library Technology Reports* 13:237 (May 1977).
5. Robert Oram, "Circulation," p.2.
6. James Robert Martin, "Automation and the Service Environment of the Circulation Manager" (Ph.D. dissertation, Florida State University, 1980), p.22.

Statistics on Headings in the MARC File

Sally H. McCALLUM and James L. GODWIN: Network Development Office, Library of Congress, Washington, D.C.

In designing an automated system, it is important to understand the characteristics of the data that will reside in the system. Work is under way in the Network Development Office of the Library of Congress (LC) that focuses on the design requirements of a nationwide authority file. In support of this work, statistics relating to headings that appear on the bibliographic records in the LC MARC II files were gathered. These statistics provide information on characteristics of headings and on the expected sizes and growth rates of various subsets of authority files. This information will assist in making decisions concerning the contents of authority files for different types of headings and the frequency of update required for the various file subsets. The National Commission on Libraries and Information Science supported this work.

Use of these statistics to assist in system design is largely system-dependent; however, some general implications are given in the last section of this paper. In general, counts were made of the number of bibliographic records, headings that appear in those records, and distinct headings that appear on the records. The statistics were broken down by year, by type of heading, and by file.

In this paper, distinct headings are those left in a file after removal of duplicates. Distinctness will not be used to imply that a heading appears only once in a source bibliographic file, although distinct headings may in fact have only a single occurrence. Thus, a file of records containing the distinct headings from a set of bibliographic records is equivalent in size to a MARC authority file of the headings in those bibliographic records.

METHODOLOGY

These statistics were derived from four MARC II bibliographic record files maintained internally at LC: books, serials, maps, and films. The files contain updated versions of all MARC records that have been distributed by LC on the books, serials, maps, and films tapes from 1969 through October 1979, and a few records that were then in the process of distribution. The files do not contain CIP records. A total of 1,336,182 bibliographic records were processed, including 1,134,069 from the books file, 90,174 from the serials file, 60,758 from the maps file, and 51,176 from the films file.

A file of special records, called access point (AP) records, was created that contains one record for the contents of each occurrence of the following fields in the bibliographic records:

Type of Heading	*Fields*
personal name	100, 700, 400, 800, 600
corporate name	110, 710, 410, 810, 610
conference name	111, 711, 411, 811, 611
topical subject	650
geographic subject	651
uniform title heading	130, 730, 830, 630

Only the 6XX subject fields that contained LC subject headings (i.e., second indicator = 0) were selected as AP records. The main entry data string was substituted for the pronoun in the series (4XX) fields that contained pronouns. The AP records also contained information from the bibliographic records that assisted in making the counts, such as the date of entry of the record on the file, the identity of the type of bibliographic file, and the language of the bibliographic record.

A third file was derived from the AP file that contained a normalized character string for each AP record heading. These normalized AP records were used to produce the counts of distinct headings by clustering like data strings. Normalization included conversion of all characters to uppercase, and masking of diacritics, marks of punctuation, and other characters that do not determine the distinctness of a heading, but would interfere with machine determination of uniqueness. The subfields included in the normalized string, hence used for all heading comparisons, are given below. Only use-dependent subfields, such as the relator subfield, and those that belonged to title clusters in author/title headings were excluded. Examples of the AP file field contents and the normalized forms are:

AP field contents:

Chuang-tzu
Chuang-Tzŭ

[Blaeu, Joan] 1596–1673
Blaeu, Joan. 1596–1673
Blaeu, Joan, 1596–1673

Byron, George Gordon Noël Byron, Baron, 1788–1824
Byron, George Gordon Nöel Byron, baron, 1788–1824
Byron, George Gordon Noel Byron, baron, 1788–1824
Byron, George Gordon Noël Byron, Baron, 1788.1824

normalized forms:

CHUANG TZU
BLAEU JOAN 1596 1673
BYRON GEORGE GORDON NOEL BYRON BARON 1788 1824

Distinct headings for this study were determined by comparing on the following subfields:

Type of Heading	*Subfields*
personal name	a, b, c, d
corporate name	a, b, k, f, p, s, g
conference name	a, q, e
topical subject	a, b, x, y, z
geographic subject	a, b, x, y, z

All occurrences of repeating subfields were included. The relator data of subfields were dropped from personal and corporate name headings as were the title subfields in author/title headings. A separate study will examine the occurrence of author/title headings. Approximately 8 percent of the name headings in the files carry title subfields: 6 percent are series and 2 percent are author/title subjects or added entries.

Two types of distinct heading counts were generated for topical and geographic subject headings. One takes account only of main terms, the *a* and *b* subfields, excluding all subject subdivisions. The other compared the complete heading strings, including subject subdivisions.

CHARACTERISTICS OF THE FILES

The four bibliographic files from which the statistics were derived were begun in different years and are of unequal size. Table 1 presents the number of bibliographic records added to each of the MARC files by the year that the record was first entered into the file. The records added in the first months of 1979 have been eliminated from tables 1–3, thus the total number of records under consideration is 1,210,809. In the combined file, the records for books dominate the contributions from other forms of materials, representing 85 percent of the combined file records. After the addition of the films and serials records in 1972 and 1973, the total number of records added each year leveled off to around 115,000 but jumped to an average of slightly more than 150,000 records per year following the ad-

Table 1. Number of Records Added to Each File by Year

Year Entered	*Book*	*Serial*	*Map*	*Film*	*Total*
1968	11,812	0	0	0	11,812
1969	43,874	0	1,104	0	44,978
1970	86,004	0	3,467	0	89,978
1971	105,390	0	8,857	6,280	114,247
1972	73,437	0	4,665	6,280	84,382
1973	92,512	3,720	5,566	8,929	110,727
1974	99,004	10,682	6,246	8,457	124,389
1975	86,527	15,866	6,721	8,604	117,718
1976	120,106	19,098	6,876	5,432	151,512
1977	140,011	17,999	7,011	4,797	169,818
1978	169,044	12,643	5,584	4,464	191,735
Total	1,027,721	80,008	56,117	46,963	1,210,809

Table 2. Numbers of Headings and Distinct Name Headings Added to All Files by Year

	Number of Headings			*Number of Distinct Headings*		
Year Entered	*Personal Names*	*Corporate Names*	*Conference Names*	*Personal Names*	*Corporate Names*	*Conference Names*
1968	14,526	3,138	155	12,620	2,139	143
1969	53,134	21,206	1,027	39,184	9,364	909
1970	104,365	42,798	2,175	63,037	14,286	1,769
1971	129,617	57,496	2,742	64,029	15,216	2,158
1972	91,040	45,768	1,942	41,246	9,891	1,402
1973	118,188	57,847	2,625	48,703	12,653	1,862
1974	127,588	73,303	2,972	51,623	17,129	1,983
1975	113,622	76,417	2,519	50,291	18,135	1,742
1976	154,718	88,207	3,454	73,182	23,120	2,306
1977	182,860	87,985	3,487	89,353	23,906	2,333
1978	218,535	97,042	4,192	99,780	24,280	2,831
Total	1,308,193	651,207	27,290	633,048	170,119	19,438

Table 3. Numbers of Subject Headings and Distinct Subject Headings Added to All Files by Year

			Number of Distinct Headings			
	Number of Headings		*First Terms Only*		*Full Headings*	
Year Entered	*Topical Subjects*	*Geographic Subjects*	*Topical Subjects*	*Geographic Subjects*	*Topical Subjects*	*Geographic Subjects*
1968	10,615	1,857	4,390	489	7,775	1,512
1969	45,161	9,047	8,104	1,980	23,617	5,426
1970	89,304	21,054	8,170	4,263	34,526	10,179
1971	115,220	31,278	6,853	5,417	36,689	12,862
1972	92,247	20,760	4,236	2,597	26,201	7,074
1973	121,161	27,890	4,460	3,105	33,061	9,819
1974	137,843	31,814	4,524	3,553	39,262	11,413
1975	130,980	30,650	4,203	3,417	40,129	11,818
1976	168,840	39,886	5,125	4,142	55,468	15,472
1977	185,331	44,973	5,718	4,194	59,529	16,676
1978	222,565	49,923	7,151	4,034	69,856	17,855
Total	1,319,267	309,132	62,934	37,191	426,113	120,106

dition of major non-English roman alphabet language records in 1976. The increase is noticeable primarily in the books and serials files since the maps file had been adding those languages since 1969 and only a limited number of non-English-language audiovisual materials are cataloged. The unusually large number of records added to the books file in 1971 resulted from a special project to add retrospective titles to the file. The large increase in books records in 1978 was due to the COMARC project in which retrospective LC records that had been converted to machine-readable form by other libraries were contributed to the LC MARC file. Approximately 12,000 COMARC records were added in 1977 and 28,000 in 1978. The fall in numbers of film records produced in 1976–1978 reflects a general fall in production of instructional films in the United States.

Counts of items cataloged that are compiled by LC processing services from catalogers' statistics sheets show that LC cataloged approximately 225,000 titles in 1978; thus, approximately 73 percent of LC cataloging is currently going into machine-readable form. The principal exclusions are records for most nonroman material (only nonroman records for maps have been transliterated and added since 1969) and a few records for music, sound recordings, incunabula, and microforms. The portion being put into machine-readable form should rise significantly as the romanized records for items in several nonroman alphabets are added in the next year.

NAME HEADINGS

Table 2 presents the number of occurrences of name headings in the MARC bibliographic files and the number of distinct name headings, both by type of heading and by year. The number of distinct headings that were new to the file in a year was determined by comparing the headings added in a given year against those added in all previous years. It is not surprising to find that 66 percent of name-heading occurrences are personal names, 33 percent are corporate, and only 1.4 percent are conference. The figures shift when considering the distinct names, where 77 percent are personal and only 21 percent are corporate.

Looking at the total figures in table 2, while 1,308,193 of the headings that appeared on the records were personal names, only 633,048 or 48 percent of these were distinct. Of the rest, 52 percent were duplicates of the distinct headings. Similarly, 26 percent of corporate names were distinct, with 74 percent being duplicates; and 71 percent of conference names were distinct, with only 29 percent being duplicates.

In 1968, 87 percent and 68 percent of personal and corporate names, respectively, were distinct, i.e., 13 percent and 32 percent "had been used previously" when they appeared on a bibliographic record during the year. As the base file of names grows, the percentage of names appearing on new records but which "had been used previously" rises, to 60 percent and 77 percent in 1974. While the figures reported in table 2 indicate that the percentage of headings used that were repeats fell slightly again in 1977 (51 percent and 73 percent), this is probably due to the influx of new names with the addition of new languages in 1976–77. Additional statistics gathered on English-language items show the percentage of repeating headings becoming steady after 1974.

SUBJECT HEADINGS

Statistics concerning distinct topical and geographical subject headings were collected for main terms, excluding subdivisions, and for full subject heading strings. Table 3 gives the numbers of headings and the numbers of distinct headings of each type found in the MARC file. Looking at the total figures, only 4.8 percent of topical first terms are distinct, the rest are duplicates. This indicates an average occurrence of 20.8 times for each first term. Slightly more, 12 percent, of the geographic first terms are distinct.

When the full headings with topical, period, form, and geographic subdivisions are considered, the percentage of headings that are distinct rises to 32.3 percent for topical subjects and 38.8 percent for geographic subjects. Thus, 67.3 percent of topical and 61.2 percent of geographic are duplicates of existing headings. In the yearly figures, subject headings show the same tendency as name headings in that the percentages of headings that appear on new records but which "had been previously used" rises as the stock of headings increases and then levels off. Subjects were also affected by the addition of other roman alphabet languages in 1976–77 but not to a very large degree.

For all access points, name headings and full string subject headings, name headings account for 55 percent of the headings that occur in the bibliographic records, with only 45 percent attributable to topical and geographical headings. It should be noted that 12 percent of the name headings that appear on the bibliographic records are names used as subjects.

FREQUENCIES OF OCCURRENCE

Counts were also made of the frequency with which name headings occurred in the bibliographic files. Table 4 summarizes the frequency data: 66 percent of distinct personal names, 62 percent of distinct corporate names, and 84 percent of distinct conference names occur only once in the files. The percent of corporate names with single occurrences is surprisingly close to that for personal; however, the percent of names having multiple occurrences falls more slowly for corporate than for personal names. While 5.47 percent of corporate names occur ten or more times, only 1.92 percent of personal names occur ten or more times. The figures for personal names roughly correspond to those obtained by William Potter from a sample taken from the main catalog at the University of Illinois at Urbana-Champaign. That study showed 63.5 percent of personal names occurred only once.[1]

The number of occurrences of different types of headings are compared in figure 1. The bars show the numbers of personal, corporate, conference, topical, and geographic headings that appear in the bibliographic files. The shaded areas represent the number of headings that are distinct, thus the upper part of each bar represents additional occurrences of the headings from the shaded area. For personal, corporate, and conference headings a further distinction is made between distinct headings that occur only once, the crosshatched area, and those that have multiple occurrences. Thus the multiple occurrences of corporate names may be seen to come from a small

Table 4. Frequency of Occurrence of Name Headings in All Files

	Distinct Personal Names		*Distinct Corporate Names*		*Distinct Conference Names*	
Number of Occurrences	*Number*	*Percent*	*Number*	*Percent*	*Number*	*Percent*
1	456,328	65.65	116,250	62.02	18,021	83.90
2	119,681	17.22	30,185	16.10	2,049	9.54
3	46,247	6.65	11,563	6.17	587	2.73
4	23,951	3.45	6,814	3.64	289	1.35
5	13,820	1.99	4,109	2.19	163	.76
6	8,790	1.26	2,958	1.58	98	.46
7	5,827	.84	2,175	1.16	56	.26
8	4,056	.58	1,673	.89	48	.22
9	2,998	.43	1,395	.74	36	.17
10	2,153	.31	10,037	.55	18	.08
11–13	4,116	.59	2,180	1.16	44	.20
14–20	3,748	.54	2,632	1.40	41	.19
21–50	2,678	.39	2,901	1.55	23	.11
51–100	448	.06	936	.50	4	.02
101–200	149	.02	374	.20	2	.01
201–300	47	.01	109	.06	1	.00
301–400	19	.00	46	.02	0	.00
401–500	11	.00	21	.01	0	.00
501–1000	5	.00	53	.03	0	.00
1001 +	2	.00	18	.01	0	.00
Total	695,074	99.99	187,429	99.98	21,480	100.00

number of distinct corporate headings, as was indicated by the slow decrease of the multiple-heading occurrence rate (i.e., a small group of corporate names have a very large number of occurrences).

FILE GROWTH

As a bibliographic file grows and the stock of names and subjects that are contained in the associated authority file increases, the number of new-to-the-file headings that are required for the new bibliographic records would be expected to fall. Figure 2 illustrates that tendency and shows that there is a leveling off of the number of new-to-the-file headings per new bibliographic record after the bibliographic file reaches a certain size. For example, after approximately 700,000 bibliographic records are in the file, for every additional 100 bibliographic records approximately 298 name and subject headings

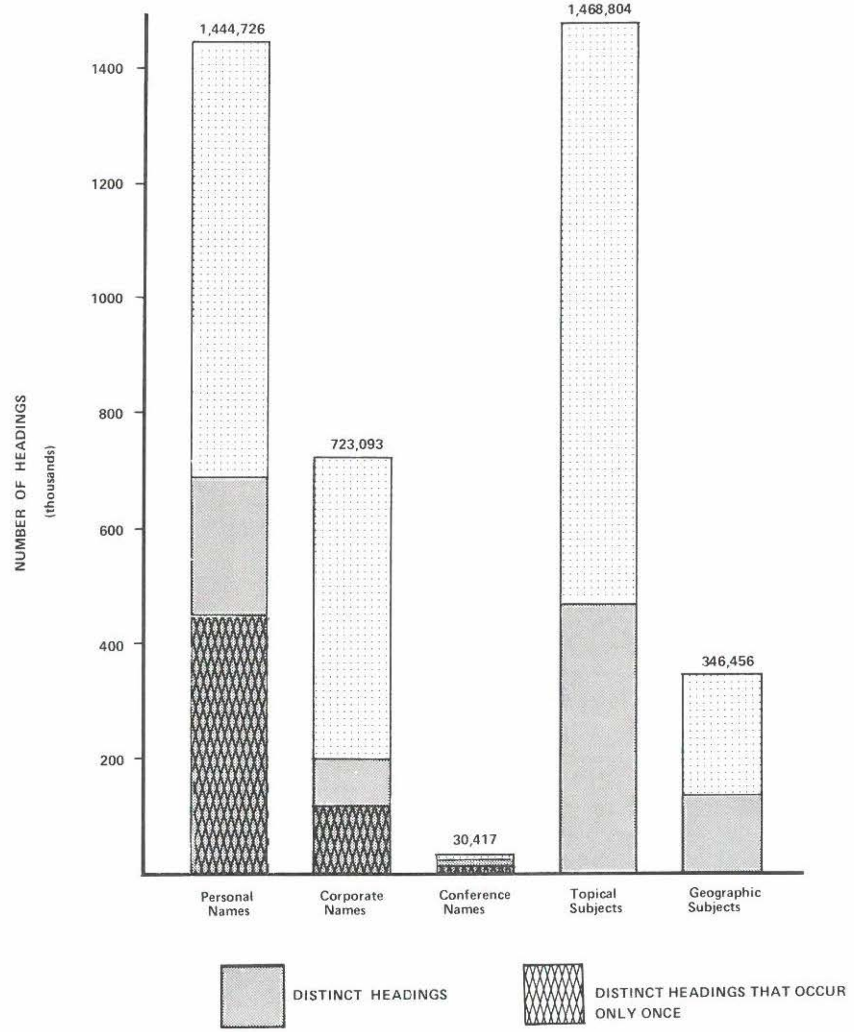

Fig. 1. Number of Headings by Type.

will be assigned, and, of these, approximately 53 will be new personal names, 14 new corporate names, 2 new conference names, 35 new topical subjects, and 10 new geographic subjects; the remaining 184 headings used will already be established in the authority file. Thus after a certain bibliographic file size is reached, the growth of the authority file is approximately a linear function of the growth of the bibliographic file.

IMPLICATIONS

The reoccurrence frequency of headings in a bibliographic file is often cited as a factor in designing bibliographic and authority-file configurations. Discussion centers on the necessity of carrying authority records for headings that occur only once in a bibliographic file. With reference to the name-heading data in table 4 and figure 1, carrying authority records only for headings that occur more than once could potentially reduce the size of the authority file from that indicated by the whole shaded area (including shaded and cross-hatched) to the plain shaded area, i.e., from 903,983 records to 310,123, a 66 percent decrease.

Controlling multiple occurrences of a heading is, however, only one role of the authority record. More important perhaps is the control of cross-references connected with the heading. Preliminary work with a

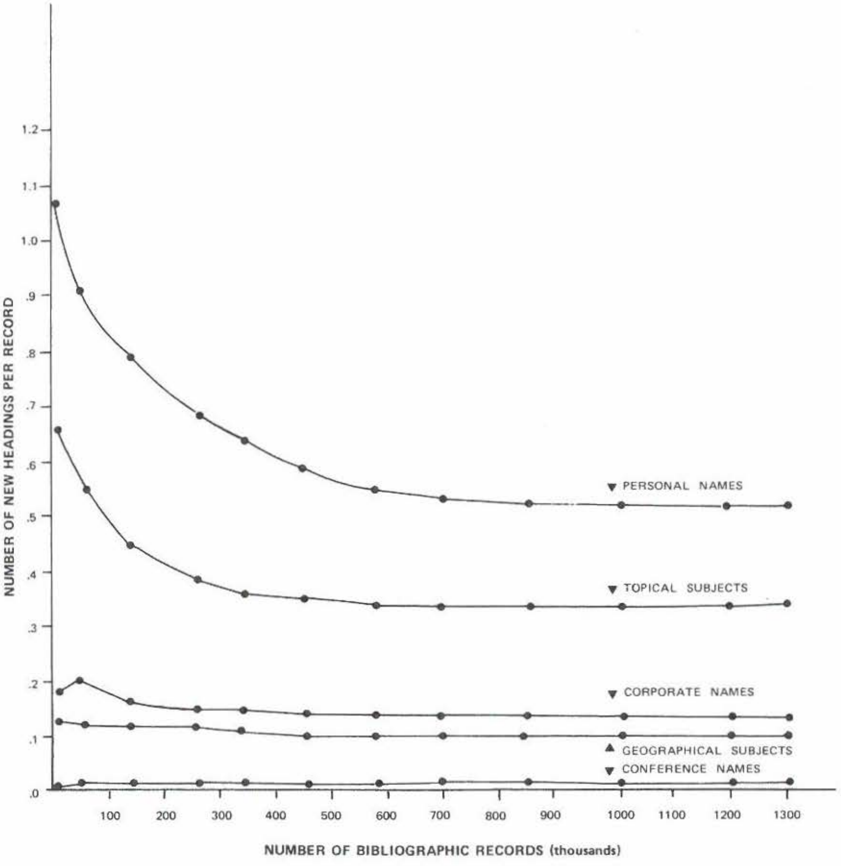

Fig. 2. Number of New Headings per Record for All Files.

random sample of personal names in the LC file indicates that less than 17 percent of personal names require cross-references. Thus the personal name headings that occur only once but would require authority records because of cross-references could be less than 17 percent. The frequency data combined with reference structure data could have a significant impact on design.

Out of a total of 695,074 personal names in the authority files associated with the MARC bibliographic files examined here, 456, 328, or 66 percent, occur only once. Of these, fewer than 77,575 would be expected to have cross-references, thus the name-authority file for personal names could be reduced in size from 695,074 records to 316,321, a 55 percent decrease. If separate authority records are a system requirement, the occurrence figures might then be useful for defining configurations that employ machine-generated provisional records for single-occurrence headings that do not have reference structures or that simplify in other ways the treatment of these headings. These figures may also be useful in making decisions on the addition of retrospective authority records to the automated files.

REFERENCE

1. William Gray Potter, "When Names Collide: Conflict in the Catalog and AACR2," *Library Resources & Technical Services* 24:7 (Winter 1980).

RLIN and OCLC as Reference Tools

Douglas JONES: University of Arizona, Tucson.

The Central Reference Department (social science, humanities, and fine arts) and the Science-Engineering Reference Department at the University of Arizona Library are currently evaluating the OCLC and RLIN systems as reference tools, to see if their use can significantly improve the effectiveness and efficiency of providing reference service. A significant number of the questions received by our librarians, and presumably by librarians elsewhere, involve incomplete or inaccurately cited references to monographs, conference proceedings, government documents, technical reports, and monographic serials. If by using a bibliographic utility a librarian can identify or verify an item not found in printed sources, then effectiveness has been improved. Once a complete and accurate description of the item is found, it is a relatively simple task to determine whether or not the library has the item, and if not, to request it through interlibrary loan.

Additionally, if the efficiency of the librarian can be improved by reducing the amount of time required to verify or identify a requested item, then the patron, the library, and, in our case, the taxpayer, have been better served. The promise of near-immediate response from a computer via an online interactive terminal system is clearly beguiling when compared to the relatively time-consuming searching required with printed sources, which frequently provide only a limited number of access points and often become available weeks, months, or even years after the items they list.

We realize, of course, that the promise of instantaneous electronic information retrieval is limited by a variety of factors, and presently we view access to RLIN and OCLC as potentially powerful adjuncts to—not replacements for—printed reference sources. Given that RLIN and OCLC have databases and software geared to known-item searches for catalog card production, our evaluation attempts to document their usefulness in reference service.

A preliminary study conducted during the spring semester of 1980–81 indicated that approximately 50 percent of the questionable citations requiring further bibliographic verification could be identified on OCLC or RLIN. The time required was typically five minutes or less. Successful verification using printed indexes to identify the same items ranged from 20 percent in the Central Reference Department to 50 percent in Science-Engineering. Time required per item averaged approximately fifteen minutes.

Based on our findings, we plan a revised and more thorough test during the fall semester of 1981–82, which will include an assessment of the enhancements to the