Toward a Computer-Generated Subject Validation File: Feasibility and Usefulness

Lois Mai Chan and Diane Vizine-Goetz

Responding to the fact that the library community has long recognized the need for improved efficiency and reliability in subject authority control, we explored the feasibility of automatically creating a subject heading validation file by scanning the OCLC Online Computer Library Center, Inc. Online Union Catalog (OLUC). The premises were, first, that although the file would not be exhaustive, it would contain the majority of frequently used headings, and second, the predicted level of accuracy in the file would be high. In approach, we focused on finding the density and distribution of assigned headings and the relationship, if any, between density and error rate. We analyzed a sample file of Library of Congress-assigned headings from the OCLC Subject Headings Corrections database. The results of the study showed that (1) frequency of use and number of headings at a given rate of use are in inverse relationship; (2) a small number of headings with high frequencies of use accounts for a majority of total use, while a large proportion shows very low frequency of use; (3) topical headings account for about two-thirds of assigned headings; and (4) error and obsolescence rates are both low, and both are in inverse relationship to the frequency of heading use. We concluded that an automatically generated subject heading file is indeed feasible. Such a file would be useful for various purposes: to verify subject heading strings constructed by catalogers, to update subject headings in catalog maintenance, and to validate subject headings during retrospective conversion.

Because subject heading validation requires extensive manual effort, subject authority control has long been one of the most labor-intensive and costly operations in library cataloging. Automatic error detection and correction mechanisms developed by OCLC Online Computer Library Center, Inc., have already reduced the need for some manual corrections, mainly those involving predictable and mechanical errors, such as spelling. But there are many types of errors these mechanisms do not catch. A subject validation file—that is, a list of valid headings and heading

Lois Mai Chan (loischan@ukcc.uky.edu) is Professor, School of Library and Information Science, College of Communications and Information Studies, University of Kentucky, Lexington. Diane Vizine-Goetz (vizine@oclc.org) is Consulting Research Scientist, Office of Research, OCLC, Dublin, Ohio. This project was supported by a research grant from OCLC Online Computer Library Center, Inc. Manuscript received for publication September 22, 1997; accepted for publication November 14, 1997.

strings that exist in a large catalog database such as the OCLC Online Union Catalog (OLUC) or the Library of Congress MARC (MAchine Readable Cataloging) database (LC MARC)—would make subject authority control much more efficient.

In former years, when the Library of Congress (LC) subject headings system was largely enumerative, *Library of Congress Subject Headings* (*LCSH*) served as a fairly effective subject validation file in spite of the fact that certain categories of headings were designated "nonprinted" and so did not appear in the list. Today, however, mainly because of how far the system has moved toward synthesis, listings in *LCSH* account for only a small percentage of the heading strings actually assigned to bibliographic records. Currently, many libraries rely on the online or print version of *LCSH* for subject authority control. *LCSH* is an indispensable tool, but its usefulness is limited by its nonenumerative nature. Accordingly, many in the profession have long felt the need for a more nearly complete subject validation file. In fact, one of the major recommendations from the Subject Subdivisions Conference held in 1991 called for the expansion of the LC subject authority file to include subject heading strings not currently listed in *LCSH* (Recommendations 1992), and three years later the Subject Authority File Task Group of the Cooperative Cataloging Council also recommended the creation of a subject validation file (Cooperative 1994).

An ideal subject validation file would contain all properly formulated subject headings in current use—in other words, it would be exhaustive and error-free. One method for developing such a file would be first to list all assigned subject headings appearing on bibliographic records and then to correct all the errors such a list would necessarily contain. Even if done centrally, the effort required would be prohibitively large at the initial stage and very high for maintenance. One must conclude, therefore, that a complete list of subject heading strings without errors is an impractical ideal. An alternative to an exhaustive subject validation file with many errors might be a smaller but relatively error-free file made up of unique heading strings that have been used frequently and account for a large proportion of usage.

Hypotheses

For such a tool to be possible, however, several assumptions pertaining to headings would have to hold. These are:

1. that frequency of use varies among headings, with some headings used over and over, and some very seldom;
2. that errors congregate at the low ranges of frequency of use; and
3. that a point exists at which errors occur at an unacceptable rate.

If investigation proved these assumptions valid, then the removal of headings below the point of unacceptability should leave a highly accurate subject validation file that could prove a useful and cost-effective professional tool.

Accordingly, OCLC's Office of Research funded a project to explore the feasibility of automatically generating a relatively error-free subject validation file that would contain all headings that had been used more than a very few times. We would examine, first, the distribution of assigned headings based on frequency of occurrence, and second, the relationship, if any, between frequency of use and error rate.

Methods

Data Collection

The Subject Heading Corrections database, developed to correct subject heading errors in the OLUC, was the source of headings for this project. The database contains an entry for each unique, complete subject heading used in bibliographic records loaded into the OCLC system through November 1992 (more than 4 million headings). A 1% sample of the headings assigned by LC was extracted from the database for further processing and examination. The sample came to 20,473 headings.

The records in the test file were ana-

TABLE 1
SAMPLE DATABASE
HEADINGS AND POSTINGS IN DESCENDING ORDER BY FREQUENCY OF USE

	Headings				Postings			
Frequency	Headings Count	Cum.Ct	%	Cum. %	Postings Count	Cum. Ct.	%	Cum. %
> 500	3	3	0.01	0.01	1,730	1,730	2.86	2.86
401–500	3	6	0.01	0.03	1,257	2,987	2.08	4.94
301–400	4	10	0.02	0.05	1,264	4,251	2.09	7.03
201–300	10	20	0.05	0.10	2,356	6,607	3.90	10.93
101–200	28	48	0.14	0.23	3,771	10,378	6.24	17.17
51–100	74	122	0.36	0.60	5,215	15,593	8.63	25.80
46– 50	17	139	0.08	0.68	810	16,403	1.34	27.14
41– 45	22	161	0.11	0.79	956	17,359	1.58	28.72
36– 40	28	189	0.14	0.92	1,056	18,415	1.75	30.47
31– 35	35	224	0.17	1.09	1,171	19,586	1.94	32.41
26– 30	47	271	0.23	1.32	1,310	20,896	2.17	34.58
21– 25	82	353	0.40	1.72	1,854	22,750	3.07	37.65
16– 20	115	468	0.56	2.29	2,058	24,808	3.41	41.06
11– 15	246	714	1.20	3.49	3,095	27,903	5.12	46.18
6– 10	792	1,506	3.87	7.36	5,959	33,862	9.86	56.04
3– 5	1,876	3,382	9.16	16.52	6,874	40,736	11.38	67.42
2	2,588	5,970	12.64	29.16	5,176	45,912	8.57	75.99
1	14,503	20,473	70.84	100.00	14,503	60,415	24.01	100.00
							100.00	

lyzed on two main parameters, distribution of headings by frequency of use and heading errors at various frequencies. The crux of the investigation was mapping error rates against the frequency distribution; also, to gain additional relevant information, supplemental analyses were carried out both before and after this mapping.

DATA ANALYSIS

The density and distribution of the sample of LC-assigned headings were determined by statistical analysis. The first step in the investigation was to determine, for each heading, how many times it appeared in the sample; then the headings were divided into groups according to frequency of use. The number of postings, or occurrences, for each frequency group—that is, the total frequency of assignment of all the headings in the group—was also tabulated. The results are shown in table 1.

The left side of table 1, labeled "Headings," shows the density of sample headings by categories of frequency of use in descending order. The columns labeled "Headings count" and "%" give the number and percentage of headings in each category of frequency ranging from 500 or greater to 1. Three of the headings in the sample file have been assigned to bibliographic records in the OCLC database more than 500 times, and more than 14,500 have been assigned only 1 time each. The heading count shows an inverse relation to frequency of use; the lower the frequency, the greater the number of headings. The most frequently assigned headings from the sample file are **Art, American—Exhibitions** (assigned 611 times), **India—Politics and government—1919–1947** (assigned 567 times), and **Family—Religious life** (assigned 552 times).

The right side of table 1, labeled "Postings," shows the distribution of postings calculated by multiplying the number of headings by frequency. A small number of

TABLE 2
DISTRIBUTION OF HEADINGS BY MARC TAG

MARC Tag	Number of Headings	Percentage	Cum. Freq.
600	2,356	11.51	2,356
610	1,145	5.59	3,501
611	26	0.13	3,527
630	156	0.76	3,683
650	13,653	66.69	17,336
651	3,137	15.32	20,473

headings account for a high percentage of use. Compare the top 122 headings (headings assigned more than 50 times) that account for about 25% of total heading usage with the more than 14,500 headings assigned 1 time each that account for approximately 25% of usage at the lower frequency levels.

The distribution of the sample of unique LC-assigned headings by MARC tag was also determined. The results, given in table 2, show that the largest number of headings are topical headings (MARC 650), followed in descending order by geographic name headings (MARC 651), personal name headings (MARC 600), corporate name headings (MARC 610), uniform titles (MARC 630), and headings for meetings (MARC 611).

To determine the extent of overlap between headings appearing in *LCSH* and the subject heading strings assigned to bibliographic records, a count was made of the number of assigned headings with a frequency of 2 or higher that matched exactly those found in *LCSH*.

EVALUATION OF HEADINGS

Details of the methods used in evaluating the sample headings are given in a previously published paper (Chan and Vizine-Goetz 1997) and are briefly summarized here.

The file of sample headings was ordered by MARC tag, frequency of occurrence of the headings, and the heading strings sorted alphabetically within frequency. The entry for each heading also included an OCLC number for a bibliographic record in which the heading appeared, the most recent date from the MARC fixed field element, "date entered on file," and a sequential ID number. The following example shows these elements for the heading **Art, American—Exhibitions.**

MARC Tag:	650
Frequency:	611
Heading string:	Art, American $x Exhibitions
OCLC#:	26855900
Date:	1992
ID number:	1589967

In the second stage of the project, we tested the relationship, if any, between the frequency of use and error rate. The heart of the investigation was to test for accuracy and to calculate error rates at various levels of frequency of assignment. In order to determine the validity of the headings, each heading had to be evaluated according to the terminology established in *LCSH* or the name authority file and the policies for combining elements in a heading. Because of the large size of the sample, it was decided to focus first on headings with higher frequencies of use. A preliminary analysis of those headings used 3 or more times each (a total of 3,382 headings), confirmed the supposition that errors indeed increased as heading use declined. Subsequently, 2,588 headings used twice each were added. This brought the working sample to 5,970. Time and labor constraints allowed examination of only a subset of the headings that were used once each.

Each of the 5,970 headings in the working sample was examined manually for correct MARC tagging, terminology, syntax, spelling, punctuation, capitalization, etc., according to the standards and

Example 1

630 Bible. $p N.T. $p Epistles of Paul $x Criticism, interpretation, etc. $x History $y Early church, ca. 30–600

Element in heading	Authority tools
630, $p, $p, $x, $x, $y	USMARC formats
Bible. $p N.T. $p Epistles of Paul	Name authority file
Criticism, interpretation, etc. \	/ *Free-Floating Subdivisions*
History >	< *Subject Cataloging Manual*
Early church, ca. 30–600 /	\

Example 2

650 Happiness $x Religious life $x Christianity $x Sermons

Element in heading	Authority tools
650, $x, $x, $x	USMARC formats
Happiness	*LCSH*
Religious life	*LCSH*
Christianity	*Free-Floating Subdivisions*
Sermons	*Subject Cataloging Manual*

Figure 1. Headings with Verification Procedure.

authority files in the following list:

- USMARC formats for authority data and for bibliographic data
- *LCSH* [both the print version and the LCXR (SUBJECTS) file in LOCIS (Library of Congress Information System)]
- Authority records in the name authority file in LOCIS
- *Free-Floating Subdivisions: An Alphabetical Index* (Library of Congress 1989–)
- *Subject Cataloging Manual: Subject Headings* (Library of Congress 1984–)
- *Revised Library of Congress Subject Headings* (Library of Congress 1991)
- *Anglo-American Cataloguing Rules*, 2d edition, 1988 revision (*AACR2R* 1988)

In a typical case, a given heading was first checked to see whether the field tag and subfield codes conformed to USMARC formats. The heading was then checked in *LCSH* or the name authority file, depending on whether it was a topical or name heading. A heading was considered valid at this point if no mechanical errors were found in coding, punctuation, capitalization, spacing, etc., and the entire heading or string, including subdivisions, matched one in the name authority file, in LCXR, or in the 16th edition (1993) of *LCSH*. The 16th edition of *LCSH* was used as the cutoff point because the cutoff date of the sample was November 1992.

The remaining headings were further evaluated with the tools listed above and analyzed in consultation with the Cataloging Policy and Support Office staff at LC.

Two examples of verification procedures are shown here. In the first case, the heading is valid. In the second case, the subdivision **$x Religious life** was incorrectly applied. The appropriate subdivision under a topical main heading is **$x Religious aspects** (see figure 1).

Categorization of Invalid Headings

Headings can be invalid because they contain actual errors or because they are obsolete in whole or in part. Through references from old to new forms, obsolescence is less detrimental to retrieval than outright error; it was thus deemed important to distinguish the two. Invalid headings were characterized as either "incorrect" or "obsolete," with headings containing both incorrect and obsolete elements classed with the former. All invalid headings were also sorted according to heading type as described in the earlier paper (Chan and Vizine-Goetz 1997). Variations within each category and examples are also given in the earlier paper.

In the spring of 1995 when the data were being analyzed, each heading that was identified as incorrect or obsolete was checked in the LOCI (the bibliographic file in the LOCIS database) in the LC

MARC database to determine whether the error or obsolete element had been corrected or updated since the sample was collected. This was done for the purpose of determining the extent of maintenance of assigned subject headings in the LC MARC bibliographic database.

After the analysis of headings with a frequency of 2 or higher was completed, a subset of headings with a frequency of use of 1, consisting of 3,472 headings (representing 23.93% of the total sample of 14,503 headings) was examined and analyzed as a test of the validity of the findings from the earlier sample. Errors in the subset were also characterized as incorrect or obsolete, and their numbers extrapolated as an estimate of the situation in the 14,503 frequency-1 headings in the full test sample.

Results

Two files figure in the summary that follows. One was the full test file of 20,473 headings, i.e., the 1% random sample of LC-assigned headings in OCLC's Subject Headings Correction database. The other, called the working file, was made up of the 5,970 headings in that file that had been used more than once. The cutoff date for both was November 1992.

The analysis of heading and posting relationships, presented in tables 1 and 2, is based on the full test file. The analyses of invalid headings, presented in tables 3–10, are based on the working file.

Distribution of Sample Headings

The data shown in table 1 support the first hypothesis that frequency of use varies among headings, with a small number of headings accounting for a large percentage of usage. Headings that were used only once or twice account for the largest number. The heading count shows an inverse relation to frequency of use; the lower the frequency, the greater the number of headings. The higher frequency categories are sparsely populated, and there is greater density in the categories of low frequencies.

The inverse relationship between heading count and frequency of use is illustrated by the following:

- 3 headings were assigned more than 500 times each, represent only 0.01% of the sample, but account for 2.86% of total usage.
- 122 headings were assigned more than 50 times each, represent only 0.6% of the sample, but account for 25.81% of total usage.
- 714 headings were assigned more than 10 times each, represent only 3.49% of the sample, but account for 46.19% of total usage.
- 1,506 headings were assigned more than 5 times each, represent 7.36% of the sample, but account for 56.05% of total usage.
- 3,382 headings were assigned 3 or more times each, represent 16.52% of the sample, but account for 67.43% of total usage.
- 5,970 headings were assigned 2 or more times each, represent 29.16% of the sample, but account for 75.99% of total usage.
- In contrast, 14,503 headings (almost 71% of the total number of assigned headings in the sample) were used only once each. However, they account for only 24.01% of the total usage.

Topical headings made up two-thirds of the full test file and 71% of the working file, with personal and geographic name headings sharing most of the remainder in each file (see table 2).

Overlap between *LCSH* and LC-Assigned Subject Heading Strings

Table 3 shows the overlap between *LCSH* and the sample headings in the working file. Only 777—just over 13%—appear in *LCSH* as such. Of these, topical headings show the greatest degree of overlap (13.94%), with personal name headings coming next (12.74%). Most of the overlapped personal name headings were those for individual families, which are enumerated in *LCSH*. The low rate of overlap between *LCSH* and name headings was due to the fact that most name headings, especially those for persons, corporate bodies, and jurisdictions, are maintained in the name authority file.

The data indicate that in assigning subject headings, catalogers seldom derive

TABLE 3
OVERLAP BETWEEN ASSIGNED HEADINGS AND LCSH

Type of Headings	Number of Sample Headings*	Headings Matching LCSH	%
600	463	59	12.74
610	181	10	5.52
611	0	0	0.00
630	39	4	10.26
650	4,233	590	13.94
651	1,054	114	10.82
Total	**5,970**	**777**	**13.02**

* Headings with frequency of use of 2 and above

headings exclusively from *LCSH*. Most name headings are based on name authority records, and over 85% of the topical headings must be synthesized.

INVALID HEADINGS

When the validation process was completed, it was found that 294 (4.92% of the 5,970 headings that were tested) were invalid, as shown in table 4. Headings considered invalid were classed as "incorrect" or "obsolete" with validity ratios figured according to the sum of the two. Among the 294 invalid headings, 76 (1.27% of the total working sample) were incorrect and 218 (3.65%) were obsolete. Thus, among the invalid headings, incorrect headings accounted for approximately one-fourth of the total, and about three-quarters were obsolete headings.

Analyzed by type of heading, uniform titles and corporate name headings showed the highest rate of invalid headings, at 17.95% and 13.26% respectively. These were followed by geographic name headings (8.35%), personal name headings (4.97%), and topical headings (3.59%). There were no headings for meetings in the sample. It might be important that over 70% of the headings in the working file were topical headings, of which fewer than 1% were incorrect and only 2.62% were obsolete, for a total invalidity rate of 3.59%. Table 4 shows the distribution as well as data on corrections in the LC MARC database during the time the study was underway.

Table 5 contains a summary of the distribution of invalid headings by frequency of use. It was found that errors do indeed accumulate at the lower levels of frequency of assignment. Obsolete headings were distributed over a wider range of frequencies of assignment.

No incorrect headings were identified among those with a frequency of use of 11 or higher, and only 7 in the set of headings that were assigned 3 times each. There was then a substantial jump to 58 in the headings assigned twice each.

On the other hand, obsolete headings were spread among headings with high frequency of use as well as among those with low frequency. Among those with a frequency higher than 20, a heading with a posting of 258 and one with a posting of 22 were both found to be obsolete. The remaining 216 obsolete headings occurred among those with frequencies of 2–20, again with a dramatic increase at frequency 2.

INCORRECT HEADINGS

Focusing on errors, table 6 contains a summary of incorrect headings with 2 or more postings. A total of 76 headings were found to contain errors.

The largest number of errors occured among topical headings, with a total of 41. Geographic name headings contained the second largest number of errors, with a total of 27. There were 6 incorrect corporate name headings, and 2 errors were detected among uniform titles. No errors were found among personal name headings, and there were no headings for meetings in the sample with a posting greater than 1.

The data in table 6 show that errors

TABLE 4
SUMMARY OF INVALID HEADINGS BY TAG
HEADINGS WITH A FREQUENCY OF USE OF TWO OR HIGHER

Tag	Total Number of Sample Headings	Number of Incorrect Headings	Incorrect Headings (%)	Number of Obsolete Headings	Obsolete Headings (%)	Total Number of Invalid Headings	Invalid Headings (%)	Invalid Headings Corrected —1995	Invalid Headings Corrected —1995 (%)
600	463	0	0.00	23	4.97	23	4.97	6	26.09
610	181	6	3.31	18	9.94	24	13.26	6	25.00
611	0	0	0.00	0	0.00	0	0.00	0	0.00
630	39	2	5.13	5	12.82	7	17.95	3	42.86
650	4,233	41	0.97	111	2.62	152	3.59	48	31.58
651	1,054	27	2.56	61	5.79	88	8.35	43	48.86
Total	**5,970**	**76**	**1.27**	**218**	**3.65**	**294**	**4.92**	**106**	**36.05**

congregated among the headings with lower frequencies of use. Fifty-eight of the 76 incorrect headings occurred in the set of headings with a frequency of 2.

Table 7 shows the cumulative ratio of incorrect headings at each frequency of use. The first four columns present the figures relating to the sample database derived from table 1. The fifth column gives the number of incorrect headings at each frequency of use. The sixth and seventh columns show the cumulative incor-

TABLE 5
SUMMARY OF INVALID HEADINGS BY FREQUENCY OF USE

Frequency	Number of Sample Headings	Number of Incorrect Headings	Number of Obsolete Headings	Number of Invalid Headings	Invalid Headings Corrected—1995
≥ 21	353		2	2	
20	28		1	1	1
19	12				
18	22		1	1	1
17	26		2	2	
16	27		1	1	
15	28				
14	36		2	2	1
13	53		1	1	
12	63				
11	66				
10	96	1	2	3	2
9	124	1	2	3	2
8	138	2	8	10	2
7	175		4	4	2
6	259	1	5	6	4
5	352	2	14	16	7
4	542	4	15	19	7
3	982	7	34	41	20
2	2,588	58	124	182	57
Total	**5,970**	**76**	**218**	**294**	**106**

TABLE 6
SUMMARY OF INCORRECT HEADINGS

Frequency	Number of Sample Headings	600	610	611	630	650	651	Number of Incorrect Headings	Errors Corrected —1995
≥ 21	353								
20	28								
19	12								
18	22								
17	26								
16	27								
15	28								
14	36								
13	53								
12	63								
11	66								
10	96					1		1	
9	124						1	1	1
8	138		1			1		2	1
7	175								
6	259						1	1	1
5	352					1	1	2	1
4	542					3	1	4	1
3	982		1			4	2	7	2
2	2,588		4		2	31	21	58	12
Total	**5,970**	**0**	**6**	**0**	**2**	**41**	**27**	**76**	**19**

rect headings count and percentage. For example, headings with a frequency of 6 and above contained a total of 5 (0.33%) incorrect headings, and those with a frequency of 3 and above contained a total of 18 (0.53%) incorrect headings. The last column is the inverse of column 7 and shows the accuracy rate; in other words, headings with a frequency of 6 or above have an accuracy rate of 99.67%, and those with a frequency of 3 or above have an accuracy rate of 99.47%.

The data also show that, except for a slight variation at the frequency of 8, the cumulative error rate increased steadily from 0% to 1.27% as the frequency of use decreased from 15 to 2.

The follow-up analysis of the subset of headings with a frequency of use of 1 showed a dramatic increase in the cumulative error rate, at 3.24% for headings in all frequency ranges; in other words, an accuracy rate of 96.76%.

OBSOLETE HEADINGS

Table 8 contains a summary of obsolete headings among headings with 2 or more postings. A total of 218 headings were found to be obsolete at the time the test database was generated.

The largest number of obsolete elements occurred among topical headings, with a total of 111. Geographic name headings contained the second largest number of errors, with a total of 61. There were 23 obsolete headings among personal name headings, and corporate name headings contained 18 obsolete headings. Five obsolete headings were found among uniform titles. There were no headings for meetings with a posting greater than 1.

The data in table 8 show that obsolete headings also congregated among the headings with lower frequencies of use. There were only 3 obsolete headings among headings with a frequency of 20 or above, while 124 of the 218 obsolete head-

TABLE 7
RATIO OF INCORRECT HEADINGS

Frequency	# of Sample Headings	Total Sample Headings— Cum. Ct.	Total Sample Headings— Cum. %	# of Incorrect Headings	Total Incorrect Headings— Cum. Ct.	Total Incorrect Headings— Cum. %	Accuracy Rate
> 25	271	271	1.32				
21–25	82	353	1.72				
16–20	115	468	2.29				
11–15	246	714	3.49			0.00	100.00
10	96	810	3.96	1	1	0.12	99.88
9	124	934	4.56	1	2	0.21	99.79
8	138	1,072	5.24	2	4	0.37	99.63
7	175	1,247	6.09		4	0.32	99.68
6	259	1,506	7.36	1	5	0.33	99.67
5	352	1,858	9.08	2	7	0.38	99.62
4	542	2,400	11.72	4	11	0.46	99.54
3	982	3,382	16.52	7	18	0.53	99.47
2	2,588	5,970	29.16	58	76	1.27	98.73

ings occurred among headings with a frequency of 2.

Table 9 shows the cumulative ratio of obsolete headings at each frequency of use. The first four columns present the figures relating to the sample database derived from table 1. The fifth column gives the number of obsolete headings at each frequency of use. The sixth and seventh columns show the cumulative obsolete headings count and percentage. For example, headings with a frequency of 6 and above contained a total of 31 (2.06%) obsolete headings, and those with a frequency of 3 and above contained a total of 94 (2.78%) obsolete headings. The last column is the inverse of column 7, showing the currency rate. In other words, headings with a frequency of 6 and above had a currency rate of 97.94%, and headings with a frequency of 3 or above showed a currency rate of 97.22%.

The data also show that, with slight variations, the cumulative obsolescence rate increased steadily from 0.37% to 3.65% as the frequency of use decreased from greater than 25 to 2.

The follow-up analysis of the subset of headings with a frequency of use of 1 showed a dramatic increase in the cumulative obsolescence rate, at 9.82%, for headings in all frequency ranges—in other words, a currency rate of 90.18%.

VALIDITY RATIOS

As shown in table 10, the validity ratio of a file made up of headings assigned more than 25 times each could be expected to be 99.63%. This ratio drops slowly as headings with lower frequencies of assignment are added, to 97.04% for a file of headings assigned four or more times each. Adding headings assigned three times each brings the ratio to 96.69%, and adding those assigned twice each brings it to 95.08%. An estimate (Chan and Vizine-Goetz 1997), based on the sample of 5,970 headings studied in this project plus a subset consisting of 3,472 headings (approximately 25%) in the frequency-1 set, indicated a cumulative invalidity rate of 9.16%, which means that the validity rate for an exhaustive file might be about 91%.

SUBSEQUENT CORRECTION AND UPDATING OF INVALID HEADINGS

In the spring of 1995, two and a half years after the cutoff date of the file from which the test sample was drawn, each heading identified as incorrect or obsolete was checked in LOCI (the bibliographic file in the LOCIS database) to determine whether it had been corrected or updated.

TABLE 8
SUMMARY OF OBSOLETE HEADINGS

Frequency	Total # of Sample Headings	MARC 600	MARC 610	MARC 611	MARC 630	MARC 650	MARC 651	Total # of Obsolete Headings	Obsolete Headings Updated—1995
≥ 259	11								
258	1						1	1	
31–257	212								
30	3								
29	16								
28	9								
27	10								
26	9								
25	10								
24	10								
23	20								
22	22	1						1	
21	20								
20	28						1	1	1
19	12								
18	22						1	1	1
17	26				1	1		2	
16	27					1		1	
15	28								
14	36					2		2	1
13	53	1						1	
12	63								
11	66								
10	96					1	1	2	2
9	124					1	1	2	1
8	138	1				3	4	8	1
7	175					2	2	4	2
6	259		1			2	2	5	3
5	352	1				10	3	14	6
4	542	1	1			9	4	15	6
3	982	4	3		2	16	9	34	18
2	2,588	14	13		2	63	32	124	45
Total	**5,970**	**23**	**18**	**0**	**5**	**111**	**61**	**218**	**87**

TABLE 9
RATIO OF OBSOLETE HEADINGS

Frequency	# of Sample Headings	Total Sample Headings—Cum. Ct.	Total Sample Headings—Cum. %	# of Obsolete Headings	Total Obsolete Headings—Cum. Ct.	Total Obsolete Headings—Cum. %	Currency Rate
> 25	271	271	1.32	1	1	0.37	99.63
21–25	82	353	1.72	1	2	0.57	99.43
16–20	115	468	2.29	5	7	1.50	98.50
11–15	246	714	3.49	3	10	1.40	98.60
10	96	810	3.96	2	12	1.48	98.52
9	124	934	4.56	2	14	1.50	98.50
8	138	1,072	5.24	8	22	2.05	97.95
7	175	1,247	6.09	4	26	2.09	97.91
6	259	1,506	7.36	5	31	2.06	97.94
5	352	1,858	9.08	14	45	2.42	97.58
4	542	2,400	11.72	15	60	2.50	97.50
3	982	3,382	16.52	34	94	2.78	97.22
2	2,588	5,970	29.16	124	218	3.65	96.35

It was found that 106 of the 294 invalid headings had been corrected or updated, 19 from the incorrect list and 87 from the obsolete list. Figures on corrections along with type-of-heading and frequency of use information are given in tables 4–6 and 8.

Summary of Results

The results of this research can be summarized as follows:

1. Distribution of subject headings assigned to bibliographic records by frequency of use: frequency of use and number of headings at a given rate of use were in inverse relationship; the higher the frequency of use, the smaller the number of headings in the set, and vice versa.
2. Relationship between number of headings and total use: a small number of headings accounted for a majority of total use. A large proportion of headings show low frequency of use.
3. Distribution by type of subject headings assigned to bibliographic records: approximately two-thirds (66.69%) of headings assigned to bibliographic records were topical headings. The remaining one-third consists of name headings and uniform titles.
4. Error rate: in the sample, headings with a frequency of 11 or above showed no errors. Headings with a frequency of 2 or higher had a total error rate of 1.27%.
5. Obsolescence rate: in the sample, headings with a frequency of 21 or above had an obsolescence rate of 0.57%. Headings with a frequency of 2 or higher showed a total obsolescence rate of 3.65%.
6. Relationship between frequency of use and error rate: the cumulative error rate was in inverse relationship to frequency of use; the lower the frequency of use, the higher the error rate.
7. Relationship between frequency of use and obsolescence rate: the cumulative obsolescence rate was in inverse relationship to frequency of use; the lower the frequency of use, the higher the obsolescence rate.

Conclusion

Based on the findings, we conclude that it would be feasible to generate a subject validation file automatically with a relatively low error and obsolescence rate, which could be used to validate a majority of subject headings assigned to bibliographic records. Our results led us to draw

TABLE 10
SUMMARY

	Headings							Postings	
Frequency	Total Sample Headings Cum. Ct.	Total Sample Headings Cum. %	Total Incorrect Headings Cum. Ct.	Total Obsolete Headings Cum. Ct.	Total Invalid Headings Cum. Ct.	Total Invalid Headings Cum. %	Combined Ratio of Valid Headings	Total Postings Cum. Ct.	Total Postings Cum. %
> 25	271	1.32		1	1	0.37	99.63	20,896	34.59
21–25	353	1.72		2	2	0.57	99.43	22,750	37.66
16–20	468	2.29		7	7	1.50	98.50	24,808	41.06
11–15	714	3.49		10	10	1.40	98.60	27,903	46.19
10	810	3.96	1	12	13	1.60	98.40	28,863	47.77
9	934	4.56	2	14	16	1.71	98.29	29,979	49.62
8	1,072	5.24	4	22	26	2.43	97.57	31,083	51.45
7	1,247	6.09	4	26	30	2.41	97.59	32,308	53.48
6	1,506	7.36	5	31	36	2.39	97.61	33,862	56.05
5	1,858	9.08	7	45	52	2.80	97.20	35,622	58.96
4	2,400	11.72	11	60	71	2.96	97.04	37,790	62.55
3	3,382	16.52	18	94	112	3.31	96.69	40,736	67.43
2	5,970	29.16	76	218	294	4.92	95.08	45,912	75.99
1	20,473	100.00						60,415	100.00

several conclusions regarding the attributes of validation files of various levels of exhaustivity. In considering these conclusions, however, it should be noted that the error and obsolescence rates that were uncovered were based on subject headings extracted from bibliographic records. Filters such as the OCLC error detection and correction program could reduce the number of errors and obsolete elements in the validation file, thus raising validity ratios above those predicted by the study results. Another factor worthy of notice is that the file tested was static for the duration of the study, and that during the two and a half years of the study about a third of the invalid headings had been corrected in the LC MARC database.

Based on the data, subject validation files of various sizes reflecting different validity rates might be generated. For example:

- *99.43% validity:* such a file would total approximately 35,000 headings with frequency of use of 21 or higher, and would account for over one-third of total use in the LC MARC database. Its obsolescence rate would be 0.57%, but it would have no incorrect headings.
- *98.29% validity:* such a file would total approximately 93,000 headings that had been used 9 or more times, and would account for almost half of total use. Its error rate would be 0.2% and its obsolescence rate 1.5%.
- *96.5% validity:* such a file would contain all headings used 3 or more times and would total approximately 340,000 headings. Its error rate would be about 0.5% and its obsolescence rate less than 3%. This file would account for approximately two-thirds of total use.
- *95.1% validity:* such a file would include headings used twice or more, and would total almost 600,000 headings. Its error rate would be 1.27% and its obsolescence rate 3.65%. It would account for approximately three-quarters of total usage.
- *ca 91% validity:* if all unique subject headings extracted from the bibliographic records in the MARC database were added, invalidity would increase considerably, to approximately 9%—almost certainly too high a rate to be acceptable. Note that this validity estimate is less reliable than those

for frequency sets 2 and higher, because it is based on a sample of about a quarter of the sample headings in the frequency-1 set.

Advantages and Drawbacks

If a subject validation file were automatically generated along the lines visualized here, at whatever validity rate policy dictates, it would have many advantages over currently available sources for subject authority control.

1. The file would contain complete subject heading strings; thus, the strings would include free-floating subdivisions as well as geographic subdivisions.
2. The file would contain all types of headings, including name headings.
3. There would be no additional cost for the creation of subject heading strings in the file, because they would be by-products of subject cataloging.
4. Quality would be high. Each heading would have been constructed or verified repeatedly during the cataloging process by human effort, so inaccuracy would be minimized.
5. Minimal intellectual effort would be required for maintenance. Both the generation and maintenance of the subject validation file could be performed by the computer. Because maintenance of subject heading strings is constantly performed in the bibliographic database, the regularly regenerated subject validation file would be a dynamic file, reflecting the corrections and updates being done continuously in LC's bibliographic database.

There are several drawbacks to this approach. The file would not be exhaustive because it would not contain all the subject headings that have been assigned. It would certainly not anticipate all possible combinations. And, until their use builds up, heading strings based on newly established headings would not likely be included. On the other hand, until the usage builds up, the number of headings affected would be small. Furthermore, heading strings based on newly established headings would not likely contain obsolete elements.

Possible Implementations

A subject validation file might be generated and displayed as a machine-readable file, a CD-ROM, or in print format—or in some combination. The machine-readable file could be used in conjunction with the electronic version of *LCSH* for automatic validation or as a tool of consultation in authority control and in original or copy cataloging. A subject validation file, with complete strings but without cross references and notes, in CD-ROM or print format could vary in scope: it could contain all subject strings above a certain frequency of use or it could be a selected file, e.g., a file of the 50,000 most frequently used subject heading strings. Such a file would be virtually error free and would have few obsolete headings. Because it would represent the most frequently written-about subjects, a selected file could be particularly useful to public and small college libraries, or to undergraduate libraries.

Other possible by-products might be discipline-based subject validation files. Such files could be extracted according to the class numbers (based on *Dewey Decimal Classification* or LC *Classification* numbers) that are associated with the subject heading strings appearing in the bibliographic records. These products could be very useful not only to subject catalogers in specialized fields but as aids in online retrieval in specific subject areas.

Potential Uses of a Subject Validation File

As a supplement and complement to *LCSH*, a subject validation file containing complete strings might help facilitate cataloging, bibliographic database management and maintenance (retrospective and routine), authority file management and maintenance (Chan 1991), as well as have additional uses.

In the area of cataloging, a subject validation file would be useful both in original and copy cataloging.

Original cataloging. Because over

85% of the subject headings assigned to cataloging records do not appear in *LCSH*, the subject validation file would be a better source of ready-made heading strings. In original cataloging, subject strings in the validation file could be used as they appear, when appropriate, thus minimizing efforts required in synthesizing subject heading strings. Furthermore, these complete strings might be used:

- as models for constructing new strings
- as means for validating and verifying headings and heading strings to ensure consistency
- as means of providing patterns for synthesizing subject heading strings involving free-floating subdivisions and geographic subdivisions

Copy cataloging. In copy cataloging, the heading strings in the subject validation file could be used to verify or update the headings found in cataloging copy.

In the area of database management, a subject validation file would be useful both for retrospective conversion and for current maintenance.

Retrospective conversion. The file could be used as a means of verifying, correcting, and updating subject headings and strings in bibliographic records.

Current maintenance. The file could be used for individual bibliographic record maintenance, as an aid in error detection and correction. It could also be used for collective maintenance:

- as an aid for adjusting local headings to additions and changes made by LC
- as an aid in maintaining currency of headings in existing bibliographic records

A subject validation file could be equally useful in machine and manual validation. For libraries and agencies that have the capability of machine validation (cf. Ludy 1985), complete subject heading strings facilitate the matching of headings in bibliographic records with those in the validation file, thus reducing drastically the number of headings requiring manual validation. For libraries and agencies that must rely on manual validation, the file could serve either as a source for or model of valid headings.

In the area of authority file management and maintenance, a subject validation file would be useful at LC as a means for facilitating the creation and revision of headings and cross references. In local maintenance, the file could serve:

- as a means of achieving consistency and compatibility within the local database
- as an aid in achieving consistency and compatibility, if so desired, between locally created headings and LC subject headings
- as a means for recording and documenting conscious choices to differ from LC practice

Finally, this file might be used as an auxiliary tool for cataloging. If it were decided to establish a regularly generated subject validation file at approximately, say, a 95% validity rate, this file could serve as a guide to help catalogers avoid the 5% invalid headings such a file would be expected to contain. Probability charts, perhaps different ones for errors and obsolescence, could indicate the likelihood of error given the type of heading and the frequency of use. The figures from this study, for instance, show that uniform titles are particularly suspect, and that corporate and geographic name headings make a worse showing than personal name headings, while topical headings have a lower invalidity rate than other heading types. The figures also show that, overall, obsolete headings are almost three times as common as headings that are incorrect on some other count. Some catalog agencies, of course, might decide to accept a given error rate in the interests of efficiency. But for those that do not, probability-of-error charts could indicate to catalogers in which circumstances additional validation work would be beneficial.

A Recommendation

Members of the profession have devoted countless hours over the last several years to considerations of how to achieve cost-effective improvement in subject authority control. The recommendations of the Subject Authority File Task Group of the Cooperative Cataloging Council (Cooperative 1994) include the long-term strategy of developing automatic validation mechanisms through the creation of subject heading records and subdivision records with appropriate coding to assist in the validation of correct synthesis of heading strings. Such a tool, when fully developed, would be an ideal solution. Its implementation would require much effort and time. As an alternative or a short-term strategy for efficient and cost-effective subject authority control, the subject validation file proposed here might prove to be viable.

Attention has centered on improving and extending the coverage of existing authority files, but projected costs have been a deterrent to action. Our results show that a major improvement can be achieved, at reasonable cost, by reducing the amount of manual review in subject authority control and minimizing the effort involved in synthesizing heading strings in original cataloging and in verifying existing headings in copy cataloging. Leading groups in the profession should seriously consider whether an instrument designed along the lines suggested should be implemented and, if so, at what level of frequency and accuracy.

Works Cited

Anglo-American Cataloguing Rules. 1988. Prepared under the direction of the Joint Steering Committee for Revision of AACR, a committee of the American Library Association, the Australian Committee on Cataloguing, the British Library, the Canadian Committee on Cataloguing, the Library Association, the Library of Congress, 2d ed., 1988 revision. Eds. Michael Gorman and Paul W. Winkler. Chicago: American Library Association.

Chan, Lois Mai. 1991. Functions of a subject authority file. In *Subject authorities in the online environment: Papers from a conference program held in San Francisco, June 29, 1987*, ed. Karen Markey Drabenstott, 9–30. Chicago: American Library Association.

Chan, Lois Mai, and Diane Vizine-Goetz. 1997. Errors and obsolete elements in assigned Library of Congress Subject Headings: Implications for subject cataloging and subject authority control. *Library resources & technical services* 41: 295–322.

Cooperative Cataloging Council. Subject Authority File Task Group. 1994. Final Report.

Library of Congress. 1910/1914–. *Library of Congress subject headings*. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

Library of Congress. 1984–. *Subject cataloging manual: Subject headings*. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

Library of Congress. 1989–. *Free-floating subdivisions: An alphabetical index*. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

Library of Congress. 1991. *Revised Library of Congress subject headings: Cross-references from former to current subject headings, compiled from the online subject authority file of the Library of Congress*, 1st ed. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

Ludy, Lorene E. 1985. OSU libraries' use of Library of Congress subject authorities file. *Information technology & libraries* 4:155–60.

Recommendations from the Subject Subdivisions Conference. 1992. In *The future of subdivisions in the Library of Congress Subject Headings System: Report from the Subject Subdivisions Conference sponsored by the Library of Congress, May 9–12, 1991*, ed. Martha O'Hara Conway, 7. Washington, D.C.: Library of Congress Cataloging Distribution Service.

USMARC format for authority data, including guidelines for content designation. 1987. Prepared by Network Development and MARC Standards Office. Washington, D.C.: Cataloging Distribution Service, Library of Congress.

USMARC format for bibliographic data, including guidelines for content designation. 1988. Prepared by Network Development and MARC Standards Office. Washington, D.C.: Cataloging Distribution Service, Library of Congress.