The Proportion of NUC Pre-56 Titles Represented in OCLC WorldCat

Jeffrey Beall and Karen Kafadar

This article describes a research project that included a designed experiment and statistical analysis to sample and estimate the proportion of records in the 754 volumes of the *National Union Catalog Pre-56 Imprints* (*Mansell*) that also appear in OCLC WorldCat. The authors randomly selected a sample of records from *Mansell* and then searched the records in OCLC WorldCat. The results show that 72.2 percent of the records in *Mansell* were found in WorldCat and 27.8 percent of the sampled *Mansell* records were not (95% confidence interval [26%, 30%]). Because a significant proportion of works held by libraries is not found in OCLC WorldCat, *Mansell* remains a valuable library resource.

The *National Union Catalog Pre-1956 Imprints* (popularly called *Mansell* after its publisher) is the largest print union catalog ever published.[1] A recent article published in *American Libraries* gave the history of the compilation of the work and described its impact on libraries and bibliography.[2] In this article, the author stated:

> At the time of *Mansell's* completion in 1981, it was estimated that 80% of its entries were not duplicated by the online network catalogs. Time, of course, has moved on and online resources now are far more extensive than 23 years ago. An admittedly very limited spot check of some *Mansell* entries in WorldCat revealed that every title I searched for was available in electronic format. It would seem that the primary value of this grand publishing venture may now be the history it provides of a bygone era.[3]

The question of overlap between *Mansell*, the largest print union catalog ever assembled, and OCLC's WorldCat, the largest online union catalog, is a valid one. Can librarians safely assume that most items represented in *Mansell* are available online? The answer has serious implications for library reference services, interlibrary loan, bibliography, and historical and genealogical research. It also provides a measure of libraries' success at the retrospective conversion of manual card files to online bibliographic records.

The validity and importance of the question, how much of *Mansell* is also found in WorldCat, requires a valid, well-designed experiment and statistical analysis. Clearly, we cannot sample all the approximately 13 million *Mansell* entries for their appearance in WorldCat. Some

Jeffrey Beall is a Catalog Librarian in Auraria Library at the University of Colorado at Denver; e-mail: jeffrey.beall@cudenver.edu. Karen Kafadar is a Professor in the Department of Mathematics at the University of Colorado at Denver; e-mail: kk@math.cudenver.edu.

statistical sampling must be done. This article describes a research project that uses statistical methods to estimate the proportion of records in *Mansell* that also appear in WorldCat.

Methodology

The basic plan of the research was to randomly select a representative sample of records from the *National Union Catalog Pre-56 Imprints* and to search for the counterpart records in OCLC WorldCat. The data response for each record was binary, depending upon whether each record was found. *Mansell*, as stated earlier, is a massive work and the largest print catalog ever published. It comprises 754 volumes, including the 69 volumes of the supplement. The entries in the catalog total approximately 13 million; this figure does not include the many cross-references included in the work.

OCLC WorldCat was chosen for several reasons. First, it is the world's largest online bibliographical database and thus serves as a good point of comparison for the world's largest printed union catalog. Second, among large bibliographic databases, it is probably the most universally available. It is possible that records not encountered in WorldCat might be found in the Research Libraries Group (RLN) union catalog, but the two databases have a large overlap. This overlap exists because both databases contain the same set of Library of Congress records that are available in MARC format. The reason for this is that some research libraries have contributed their records to both utilities and two different libraries often catalog the same work in each of the two databases. Because WorldCat is larger, one would assume it is more likely that a particular record will be found there.

In designing their sampling experiment, the authors considered the possibility that the proportion of *Mansell* records found in WorldCat might vary by volume year. Therefore, the experiment design allowed for some comparisons among proportions estimated from *Mansell* volumes from different years. The volumes were first divided according to year of publication (1968 to 1981), as indicated in table 1. From the volumes in each year, the authors used the random number function in R (http://www.r-project.org), a statistical analysis software system, to select two of the approximately sixty volumes from each of the years 1969 to 1980 and one volume from each of the years 1968 and 1981 (due to the fact that these two years contained fewer volumes, 5 and 34 volumes, respectively). The authors thus selected, at random, 26 volumes from among the 754 that make up the entire set of *Mansell* volumes, distributed throughout the entire time period.

A perusal of the volumes revealed 698 pages in all volumes in the main set (i.e., nonsupplement) except volume 685, which has 874 pages. The number of pages in the supplement volumes varies slightly from

TABLE 1
Breakdown of the Individual Years of Publication for Volumes of *Mansell*

Year of Publication	Volume Numbers	Total Number of Volumes
1968	1–5	5
1969	6-64	59
1970	65–124	60
1971	125–184	60
1972	185–244	60
1973	245–304	60
1974	305–364	60
1975	365–424	60
1976	425–484	60
1977	485–544	60
1978	545–604	60
1979	605–664	60
1980	665–720	56
1981	721–754	34

volume to volume but is generally around 560. The pages in all volumes in the main set and the supplement contain three columns on each page and usually seven records per column, but ranging from 6 to 8 depending on the "size of the card" (i.e. length of the record and number of cross-references present). The authors therefore prepared a list (using the R software) for each volume of 30 (page number, column number, record number) triples, where they randomly selected "page number" from the numbers {1, 2, …, 694}, randomly selected "column number" from the numbers {1,2,3}, and randomly selected "record number" from the numbers (1,2, …, 7). They did not check records in non-Roman languages (skipping that triple) but included records in any other (Roman) language. If a column had only five records and the design required the number "7," the authors also skipped that triple to avoid the potential for oversampling of records near the tops of the pages. (The complete set of record locator triples is available upon request from the authors.)

To illustrate the sampling process, the first triple in the design was volume 3, page 41, card 1. This card was found in WorldCat, so "1" (found) was recorded as the outcome. Although 30 triples per volume were listed, statistical consideration dictated that only 20 records would probably give sufficient precision in the estimates of "proportion found." The authors tried to complete 20 searches for each of the 26 volumes. If the random specification (triple) fell on a non-Roman entry, that specification was skipped over to the next random selection. Similarly, if the search specification fell on a cross-reference, that instruction was skipped and the next specified card was searched. In a few cases, the random specification required a higher card number than number of records in the particular column to be searched; in these cases also, the authors skipped the specification and searched for the next specification. For each volume, the authors stopped searching for records when 20 results had been recorded. They did not check skipped searches and hence did not record them as either "found" (1) or "not-found" (0). For four of the 26 volumes, more than 10 searches fell on cross-references and the 30 random selections were used before 20 searches could be completed. Thus, the total number of searches was slightly less than $26 \bullet 20 = 520$ (but only by 12) and 508 records allowed more than adequate precision in the overall estimate of "proportion found."

A cataloger with more than fourteen years of experience searching OCLC WorldCat performed the searching in OCLC, looking for records in the summer of 2004. In cases where the cataloger did not find a matching record, he made several attempts to find the record online by performing alternative types of searches, such as title, author, series, alternate author, and so on. To be considered a match, the item in OCLC had to match exactly the item in *Mansell* (i.e., exactly the same edition, format, year, and so on). When a particular search called for a third edition of a work and only the first edition was found in OCLC, the search was counted as "not found." Similarly, if the search called for a print edition of a work and the only edition found in WorldCat was a microform edition, the search was counted as "not found." Although a microfilm edition of a work contains the same intellectual content as the original print edition of the work, the authors still chose to count these as "not found" when the microfilm edition was not in OCLC because the goal was to estimate the proportion of records, not to find whether a facsimile was available.

Results

The results of the 508 searches completed are listed in table 2. A total of 367 (72.2%) of the 508 items were found in WorldCat, and 141 (27.8%) were not found. A 95 percent confidence interval for this proportion is (0.70, 0.74).[4]

Statistical Analysis

The overall proportion of records found is $= 0.7224$. The authors also checked for any systematic trends over time in these proportions among volumes; a plot of "Proportion found" versus volume number (or versus 1,2,3,…,26) did not show any strong trends, except perhaps slightly higher proportions (above 85%) for five of the last nine volumes. Statistically, these "slightly higher" proportions are not significantly different from the other 21 proportions.

As shown in table 2, the four volumes with fewer than 20 records were volume numbers 15, 22, 23, and 26. Among the remaining 22 volumes that had data on all 20 records, the authors checked to ensure that the numbers found (out of 20) were consistent with the variation they would expect from a binomial distribution with $n = 20$ and $p = 0.7224$.[5] The mean of this binomial distribution is 14.45. In addition, the binomial distribution gives an expected number of 10s, 11s, 12s, …, 18s that can be compared with the observed number of 10s, 11s, 12s, …, 18s. These numbers, observed and expected, are given in table 3.

TABLE 2
Data Collection*

Search No.	Vol. Searched	No. of Records Found	No. of Records Not Found	Total
1.	3	11	9	20
2.	23	12	8	20
3.	56	16	4	20
4.	88	13	7	20
5.	108	16	4	20
6.	154	15	5	20
7.	159	13	7	20
8.	220	15	5	20
9.	224	15	5	20
10.	301	11	9	20
11.	303	16	4	20
12.	333	13	7	20
13.	350	16	4	20
14.	403	13	7	20
15.	419	10	8	18
16.	432	14	6	20
17.	463	15	5	20
18.	504	18	2	20
19.	538	17	3	20
20.	555	11	9	20
21.	590	18	2	20
22.	613	12	7	19
23.	621	18	1	19
24.	669	18	2	20
25.	671	12	8	20
26.	730	9	3	12
Total	—	367	141	508

*The columns show the search number, the volume searched, the number of corresponding records found in OCLC WorldCat, the number not found, and the total number of records searched for each volume.

A chi-squared goodness of fit statistic [sum of (observed-expected)^2/expected] = 12.27.[6] A corrected statistic, which is better for small counts, yields 12.00.[7] Neither is significant at the 0.05 level, indicating that the observed counts are consistent with the expected counts from a binomial distribution with $p = 0.72$.

Conclusion

The data from this study show that a significant percentage (27.8%) of the sample of records selected from the *National Union Catalog Pre-56 Imprints* are not represented by records in OCLC WorldCat. This finding has important implications for library reference services, interlibrary loan (ILL), bibliography, and historical and genealogical research. The study suggests that retrospective conversion is far from complete in North American libraries overall. It is not safe to assume that all materials

TABLE 3
Consistency of Proportions Across 22 Volumes*

k = Number of records found	10	11	12	13	14	15	16	17	18
Observed number	0	3	1	4	1	4	4	1	3
Expected	0.43	1.01	1.97	3.16	4.11	4.28	3.48	2.13	0.93

*Row 1 lists the number (k) of records out of 20 that can be found in each search. (Only 10 through 18 are listed because no volume produced fewer than 10 records or more than 18.) Row 2 gives the observed number of volumes whose searches yielded (k) records found. Row 3 gives the expected number of volumes according to a binomial distribution with 20 trials (searches) and p = 0.7224 probability of successfully finding a record. The chi-squared goodness of fit statistic (Σ [obs-exp]^2/exp, equals 12.27), which is not significant at p = .05, when compared with a χ^2_9.

held by North American libraries are represented in the largest online bibliographic database, WorldCat. Librarians, especially reference and ILL librarians, will benefit from this knowledge, for they will be reminded to direct patrons to search *Mansell* or to use it as a source of interlibrary loans, as a supplement to WorldCat. The data also remind researchers and librarians of the enduring value of the largest print bibliography ever published.

Notes

[1] *The National Union Catalog Pre-1956 Imprints: A Cumulative Author List Representing Library of Congress Printed Cards and Titles Reported by Other American Libraries* (London: *Mansell*, 1968–1981).

[2] Danelle Hall, "*Mansell* Revisited," *American Libraries* 35, no. 4 (Apr. 2004): 78–80.

[3] Ibid., 80

[4] George W. Snedecor and William G. Cochran, *Statistical Methods*, 6th ed. (Ames: Iowa State University Press, 1967), 5.

[5] Ibid., 210.

[6] Ibid., 21.

[7] Timothy R. C. Read and Noel A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data* (New York: Springer-Verlag, 1988), 3.