The Effectiveness of Copy Cataloging at Eliminating Typographical Errors in Shared Bibliographic Records

Jeffrey Beall and Karen Kafadar

Typographical errors in bibliographic records can cause retrieval problems in online catalogs. This study examined one hundred typographical errors in records in the OCLC WorldCat database. The local catalogs of five libraries holding the items described by the bibliographic records with typographical errors were searched to determine whether each library had corrected the errors. The study found that only 35.8 percent of the errors had been corrected. Knowledge of copy cataloging error rates can help underscore the importance of quality data in bibliographic utilities and, further, can serve as an indication to libraries whether they need to pay more attention to correcting typos in the copy cataloging process.

Copy cataloging, the process of copying bibliographic records from a source database such as OCLC WorldCat, has increased librarians' efficiency by eliminating duplication of effort. One library creates a bibliographic record for an item such as a book and many other libraries can copy or migrate the data into their local online catalogs, thus saving each individual library the work of cataloging the item and entering the data into the system.

However, the ability to copy data from other libraries potentially can detract from the value it adds to the cataloging process. Libraries that copy data from a bibliographic record in the source database can also copy typographical errors made in the record.

Libraries differ in the amount of quality control they perform during the copy cataloging process. Several factors relating to the source of the bibliographic records (such as records created by the Library of Congress) may affect the amount of editing or quality control an individual record receives. This paper describes a study that sought to answer the question, "How successful are copy catalogers at finding and correcting typographical errors found in bibliographic records imported from OCLC WorldCat?"

Previous Studies

Only a few papers have reported on the extent of typographical errors originating in cataloging copy and remaining uncorrected in local library online cata-

Jeffrey Beall (jeffrey.beall@cudenver.edu) is Catalog Librarian, Auraria Library, University of Colorado at Denver; **Karen Kafadar** (kk@math.cudenver.edu) is Professor of Mathematics, University of Colorado at Denver.

The authors wish to thank the students in the spring 2002 Statistical Consulting Workshop class at the University of Colorado at Denver for help in designing the study. The 2002 Samuel Lazerow Fellowship, funded by the Institute for Scientific Information, supported the research.

logs. A 1989 paper by Sheila Intner titled, "Quality in Bibliographic Databases: An Analysis of Member-Contributed Cataloging in OCLC and RLIN," compared the quality of data between the two utilities and found it to be similar.[1] The points of comparison included such elements as adherence to cataloging rules, tagging errors, and spelling errors. Intner did not use the term "typographical error" and refers to all such errors as spelling errors. She found that "simple spelling and tagging errors, troublesome wherever they occurred, affected retrieval negatively in headings, while errors in capitalization and punctuation usually did not, although they look peculiar."[2]

A brief report in *American Libraries* in 1991 described a rough method of determining the quality of a particular bibliographic database, suggested by Jeffrey Beall, that entailed performing keyword searches of ten misspelled words and counting how many records were retrieved in the searches.[3] The method was called the "Dirty Database Test." This report inspired and influenced several other writers who improved and followed up on the idea. Jim Dwyer described the test and the reactions it generated from catalogers on two cataloging-related electronic discussion lists.[4] He described both positive and negative reactions to the test and reported that one posting "commented that any test which might result in a cleaner data base was of some use."[5]

A paper by Terry Ballard in 1992 improved on the Dirty Database Test and described a systematic method for eliminating typographical errors from a database.[6] His method involved using a particular feature of the INNOPAC integrated library system to search through every keyword in the database (a lengthy task) and identify and correct obvious errors. The paper included a list of the most common misspellings found in Ballard's local database and invited readers to search and correct these misspelled words in their own local databases. A second paper by Ballard and Arthur Lifshin from the same year analyzes typographical errors themselves.[7] They found that "all of the words that are misspelled many times tend to have eight or more letters and at least three syllables."[8] Moreover, "it is the more common words that have been misspelled and not the more esoteric technical terms."[9] They suggest, "Every library that has an OPAC with keyword capability should search the problem words that we have identified and fix the inevitable errors."[10]

Another paper in 1992 by Sylvia Gardner examined spelling errors in online databases from a user's point of view.[11] Like other authors writing on typographical errors, she made little distinction between spelling and typographical errors. She classified the four types of typos as errors of letter omission, errors of letter insertion, errors of letter substitution, and errors of letter transposition.[12] Describing the negative impact of spelling errors on database users, Gardner claimed there was a "reduced recall and precision in the retrieval of information."[13]

In their paper, "Lost Articles: Filing Problems with Initial Articles in Databases," Ralph Nielsen and Jan M. Pyle found the "quantity of such errors . . . to be high."[14] They studied bibliographic records representing works in European languages and noted that "every single error represents a title that will not be found by someone looking for it."[15] Barbara Nichols Randall, on the other hand, in her paper, "Spelling Errors in the Database: Shadow or Substance" concluded, "Most spelling errors are redundant errors and thus do not prevent users from finding the needed record."[16] She attributed most typographical errors in the database she studied to lax standards during retrospective conversion.

A 2002 monograph by David Bade, *The Creation and Persistence of Misinformation in Shared Library Catalogs: Language and Subject Knowledge in a Technological Era*, presents a philosophy of errors in bibliographic records.[17] Bade provides many thoughtful and provocative insights on the impact of all types of errors—such as linguistic, typographical, and cataloging—in bibliographic records. He states, "Mistakes in MARC coding of bibliographic and authority records, whether as typographical mistakes or improper coding, is a greater problem since they can seriously disrupt a user's ability to find and interpret bibliographic information."[18] Referring to libraries' using copy from the bibliographic utilities, Bade claims, "If catalog records from these external sources have any inadequacies or errors, the library will be paying for, and living with a great body of misinformation."[19] He criticizes the current state of copy cataloging and its high error rate and says, "'See no evil, fix no evil,' applies to much of the copy-cataloging done in academic libraries. As a result, bad records persist and are being edited locally by each institution according to 'whatever' standards: the exact opposite of how shared databases should function."[20] He continues, "By accepting without review these various kinds of records, the quality of the shared database is undermined."[21] Bade offers a potential solution to the problem of errors in shared bibliographic records. He suggests, "Any librarian can spot the errors and report them to the appropriate person."[22]

The Importance of Studying Typos

The presence of a typographical error in a bibliographic record can adversely affect the ability of a library user to find needed information, or, in other words, "a single error can render a document virtually irretrievable."[23] Typographical errors can occur in almost any part of a bibliographic record. Errors that occur in headings, such as authors, titles, and subjects, can be more of an obstacle to library users because they may cause a particular record not

to be retrieved in an OPAC search and thereby prevent a user from accessing information about an item that the library actually holds. For example, if a user is looking for a particular work by Shakespeare and the author heading for the bibliographic record for that work uses an erroneously spelled name, "Shkespeare, William, 1564–1616," the error will prevent the user from accessing the desired work.

Typos that occur in the non-heading elements of a bibliographic record, such as contents notes, also can obstruct access when the data in these fields are included in a library's keyword indexes. If a word is misspelled in one heading in a record and then is spelled correctly elsewhere in the same record, then retrieval may not be affected. However, some library catalogs have precise keyword searching capabilities, such as specific keyword author or keyword subject searches, so a second, correctly spelled instance of a misspelled word does not always get included in a specific keyword index. Moreover, a word containing a typographical error may be the only instance of that word in the entire record.

The copy cataloging process often involves migrating or copying bibliographic records from a bibliographic utility, such as OCLC WorldCat, into a library's local integrated library system. If an error exists in a bibliographic record in a utility, then library copy catalogers have the opportunity to correct the error at the time of copy cataloging. Libraries have different policies for verifying data quality in copy cataloging. Some libraries do little or no checking of records for data quality, such as correct form of heading, accuracy in transcription of title and other data, and absence of typographical errors. Some libraries apply different levels of scrutiny depending on the source of the record. For example, a library may accept all records that originate from the Library of Congress (LC) without any editing or quality control but do a more thorough check of records from non-LC libraries, even though typographical errors can and do occur in records created by the Library of Congress.

This study was designed to characterize the degree to which cataloging departments have been successful in finding and correcting errors that occur on shared bibliographic records. Knowledge of the copy cataloging error rates helps to underscore the importance of quality data in the bibliographic utilities and further, can serve as an indication to libraries whether they need to pay more attention to correcting typos in the copy cataloging process. This study did not look at the proportion of typographical errors in a given bibliographic database in relation to the size of that database. Clearly, a large database with a thousand errors is not as serious a problem as a small database with the same number of errors. Instead, this investigation looks at how successful copy cataloging in general is at correcting typographical errors.

Scope of Typographical Errors

The typographical errors investigated in this study are those made by catalogers or those doing data entry—not the typographical errors that occur in published items that are followed by the error indicator [*sic*] in the bibliographic record. This study considers genuine typographical errors including misspellings, transposed letters, and missing letters. Only English language words were examined in the study.

The typographical errors are taken from the Web site titled, "Typographical Errors in Library Databases," which is maintained by Terry Ballard.[24] This site provides a list of the most common typos that tend to occur in online library catalogs. The authors used the list of typos as it existed in May 2002. New words are added regularly to the list and it is supplemented by an electronic discussion list, with librarians cooperatively contributing typographical errors as they encounter (or make) them. This list of common library OPAC typos is extensive and includes over a thousand words. It is divided into five categories that correspond to the probability of encountering the typo in a library database: very high, high, moderate, low, and very low.

Research Project

Study Design

The effort involved in this study dictated that a maximum of about 500 individual bibliographic records could be examined. The response for each record was binary, either "corrected" or "not corrected." This number (500) was split among word frequency categories (f), words within each category (w), and libraries to be examined for each word (n); in other words, $f \cdot w \cdot n \ ? \ 500$. With f (word frequency categories) equaling five (very low, low, medium, high, very high), the product of w (number of words in each frequency category) and n (number of libraries to query for each possibly misspelled word) was constrained to be not more than one hundred. The choice of w and n (say, $w = 4$ different words and $n = 25$ libraries for each word, or $w = 20$ and $n = 5$, or $w = 10$ and $n = 10$) involved considerations of expected variability within libraries on a given word or within words in a given word frequency category. For example, checking twenty-five libraries for each of four words would yield more precise estimates of the proportion of libraries that had corrected four specific words, but yield no information at all on other words. This strategy would be sensible if the probabilities of corrections were basically the same for all words. However, it was deemed more likely that these probabilities might vary considerably for different words. For that reason, larger values of w were selected, at the expense of having less information (smaller n) on the probability of correction for each of the words. In this study, twenty words in each cate-

gory were randomly selected from a list according to a random number table, and online library catalogs from five libraries among those that listed the record in their holdings were examined to see if the error had been corrected.[25]

To obtain a valid estimate of the error rate across multiple library catalogs and different types of words, a carefully designed study was needed. A convenience (non-random) sample of records would potentially have been inadequate for several reasons. First, a convenience sample might have resulted in the use of frequently accessed records, which are hardly representative of all records in a given library's catalog. Second, frequently accessed records could have afforded more chances for errors to be noticed and possibly corrected, so the true error rate may be underestimated if based on such a sample. Finally, more common words might have appeared to be misspelled more often simply because they appear more often. A study to estimate the overall error rate needed to take into consideration both the frequency of the misspelled words in the English language as well as their likelihood of being misspelled.

We started with a table of word frequency.[26] We also used a list of words commonly misspelled.[27] Five categories of word frequency and five categories of likelihood of misspelling were identified (very high, high, moderate, low, very low). The strong dependence between these two factors—word frequency and likelihood of misspelling—became readily apparent; words that are very common often showed up in the list of frequently misspelled words, and vice versa. Thus we abandoned the first factor, word frequency, in our stratification of words and sampled words within only five categories of likelihood of misspelling.

Gathering the Data

The basic strategy of this study was to take a random sample of errors found in OCLC bibliographic records, determine which libraries had used or copied the bibliographic record into their local systems, and then examine a sampling of those local systems to determine what proportion of the libraries had corrected the errors.

For help in designing this study, we presented it as a class project in the Mathematics Department of the University of Colorado at Denver. The class, Statistical Consulting Workshop, works on real-world statistical problems presented by members of the local community. The Math Department charges a small fee for this service, which benefits a departmental fund. To cover the fee, we used money from the 2002 Samuel Lazerow Fellowship awarded by ACRL to support this research. The class was taught during the spring semester of 2002.

Based on a recommendation from the statistical consulting class, we randomly selected twenty words from each of the five categories for a total of one hundred words. The sample size was dictated by a desire to obtain a reasonably precise estimate of the overall error rate as well as an indication of whether this error rate was consistent across the five categories. The randomization was done using tables of random numbers provided to us by the class. In selecting twenty words from each category, we chose enough words to have a reasonably stable estimate of the probability that the ratio of corrected words to uncorrected words would not greatly vary within any given category. We needed to examine enough words to rule out the possibility that the frequency category of a word did not determine whether or not it was more likely to be corrected. For each of the one hundred misspelled words, we searched the online catalogs of five libraries selected at random from a random number table. The study design, therefore, took into account that the number of corrected errors might differ according to category

Next, we performed a keyword search for the misspelled words in OCLC to find suitable records containing the errors. We performed author, title, subject, and note keyword searches to find records containing the misspelled words. Finding records containing typos from the "very high" probability category was generally easier than finding typos from the "very low" probability category. In some cases, the word itself determined what type of keyword search should be used. For example, for the typo "pictorialworks," which is the two words "pictorial" and "works" run together without a space, we did a subject keyword search because the term "pictorial works" occurs most frequently in bibliographic records as a subject form subdivision. We sought records that both contained the particular typo and had at least ten or more holdings (records that resulted in fewer than ten holdings referred to rather uncommon words).

After finding a suitable record for each misspelled word, we printed the list of holdings that corresponded to the record. Using a list of random numbers provided by the statistical consulting class, we determined the first of the five holding libraries whose catalogs we would examine. For example, if the next number on the random list was seventeen, we counted to the seventeenth library in the holdings list.

To determine the other four libraries from the holdings list, we first counted the total number of holding libraries listed and divided that number by five. Continuing the example from above, for a record that contained fifty holdings, we would divide fifty by five. The dividend, ten, would become the spacing increment between that first holding library and the other four in the list. With the first library being number seventeen on the list, we would thus also examine libraries numbered twenty seven, thirty seven, fourty seven, and seven. We would start around back at the beginning of the list whenever we ran to the end of the list of libraries. In this manner, bibliographic records from five

libraries were randomly selected and examined for each of the one hundred typographical errors.

Examining the Records

The purpose of examining the records was to determine if each randomly selected library had corrected the typo. Though this seems straightforward, this step actually turned out to be the most difficult part of the study. We soon learned that some libraries had their holdings listed in OCLC but did not migrate the OCLC record into their local system. Instead, they obtained the record from some other source. When we encountered this situation, we selected the next library in the holdings list, because the typo would not be present in the record the library used. Before we determined whether a particular library had corrected a typo, we used several methods to be very sure that the library was indeed using the same record that contained the typo. First, whenever possible, we looked at the MARC display in the local system. (More and more integrated library system (ILS) vendors include this functionality in the public mode of their online catalogs.) Upon viewing the MARC display, we compared the OCLC number in the record with the number on the master OCLC record and verified that they matched. If they did not match, we selected the next library from the list and began the process again.

In some instances, the typographical error was present in a field that had been added to the record some time after the record had been created. When we examined the records, we determined that most did not contain the field with the typo. In these cases, we eliminated the master record containing the typo and found a new record with the same typo.

The Data

We looked at the online catalogs of five randomly selected libraries for each of the one hundred typographical errors, for a total of 500 individual bibliographic records examined. We found that, out of the 500 records, 179, or 35.8 percent, had been corrected, and 321, or 64.2 percent, had not been corrected. Table 1 shows the number and percentages of errors corrected and uncorrected. A 95 percent confidence interval for the proportion of remaining errors is (60.0 percent, 68.4 percent). That is, if this same study were repeated, in exactly the same manner, one hundred times, and a 95 percent confidence interval was computed for each of those one hundred times in exactly the same manner, then ninety-five of those intervals would cover the true proportion of remaining errors.

Table 1. Numbers and percentages of errors found corrected and not corrected

	Total corrected	Total not corrected	Total
Number	179	321	500
%	35.8	64.2	100

Statistical Analysis

The misspelled words that were randomly selected for this study are listed in table 2. The overall proportion corrected in each word frequency category also is stated in this table. Consider, for example, the word "literature" in the high frequency category, misspelled as "literatue." Among the five randomly selected libraries with this holding, two of them had corrected this misspelling and three had not, resulting in an estimated probability of correction of .40 (40 percent). A fuller version of table 2 is available in Appendix 1 and includes explanations of the typographical errors and data about the MARC field in which each typo occurred.

Figure 1 displays these one hundred proportions (twenty in each word frequency category) using a box-and-whiskers display.[28] The center line in each box is located at the median in each group (in other words, the average of the tenth and eleventh largest proportions among the twenty). The lower and upper ends of the box appear at the lower and upper quartiles (that is, the average of the fifth and sixth proportions, and the average of the fifteenth and sixteenth proportions, respectively). The "whiskers" extend out to the extremes (minimum = 0 and maximum = 1 in the first two categories, 0 and 0.8 in the last three categories). Notice that for six of the twenty words in the "low" category, zero out of the five sampled libraries had corrected the record, so the lower quartile is the same as the minimum (zero).

Figure 1 and table 2 both suggest that the proportion of words corrected in the record may depend on the word frequency. This proportion seems to be about 0.40 (40 percent) if the word is in the very high (VH), high (H), or moderate (M) frequency category, but somewhat lower, about 0.30 (30 percent), if the word frequency is low (L) or very low (VL). Combining the data on the sixty words in the first three categories yields an estimated proportion corrected of 120/300 = 0.40; among the forty words in the last two (low frequency) categories, the estimated proportion corrected is 59/200 = 0.295.

To test our hypothesis that the true proportions of corrected words in these two groups is the same, we compare 0.40 and 0.295 using a conventional two-sample test of proportions.[29]

$$\frac{(0.40 \quad 0.295)}{\sqrt{\frac{(0.40)(0.6\ 0)}{300} \quad \frac{(0.295)(0.705)}{200}}} = \frac{0.105}{0.043} = 2.44$$

This is statistically different from zero at the $a = 0.05$ level of significance (two-sided p-value is 0.0146). These data suggest that the probability of correction depends on word frequency, which is not surprising. A 90 percent confidence interval for the proportion corrected in the VH + H + M categories is:

$$0.40 \pm 1.645 \times \sqrt{\frac{(0.40)(0.60)}{300}} = (0.353, 0.447) = 35.3\% \text{ to } 44.7\%$$

A 90 percent confidence interval for the proportion corrected in the L + VL categories is:

$$0.295 \pm 1.645 \times \sqrt{\frac{(0.295)(0.705)}{200}} = (0.242, 0.348) = 24.2\% \text{ to } 34.8\%$$

A 95 percent confidence interval for the difference in proportions is (0.021, 0.189) = 2.1 percent to 18.9 percent; that is, the true difference is likely (with probability 0.95) to be at least 2.1 percent and no more than 18.9 percent. A box and whisker plot of the data combined into the two groups is shown in figure 2. The "notches" in the boxes show the approximate limits of a 95 percent confidence interval for the medians of the groups.[30]

During the data collection process, it appeared as if the location of the word within the individual MARC field might affect its likelihood of being corrected. Figure 3 is a plot of the "depth" in the field (i.e., location of the word among the k words on the line) as a function of the estimated proportion of libraries that corrected the word (x-axis).

The mean depth is denoted by an enlarged "**x**" when depth is 0–0.20, 0.20–0.40, 0.40–0.60, 0.60–0.80, 0.80–1.00. The data do not suggest an association between depth and correction probability, so this hypothesis was not investigated further.

Conclusion

A random sample of records containing misspelled words and a random sample of five of the libraries whose online catalogs contained bibliographic records with these misspelled words revealed a surprisingly large proportion of records that remain uncorrected. This proportion appears to depend on the frequency of the words: very high, high, or moderate frequency words tend to be corrected about 40 percent of the time (90 percent confidence interval: 35 percent to 45 percent), while low or very low frequency words tend to be corrected only about 30 percent of the time (90 percent confidence interval: 23 percent to 36 percent). This study was not large enough to detect an effect of "depth"; in other words, an association between the location of a word in the MARC field and the total number of words in the field may affect the proportion corrected, but the data are insufficient to confirm this hypothesis.

Libraries can take several steps to eliminate typographical errors in bibliographic records and improve access. First, libraries can search their catalogs for the common typographical errors in the list created by Terry Ballard. Second, utilities and other suppliers of bibliographic records can routinely search and correct errors in their master databases. This work can be done by professionals on the utilities' staffs, but OCLC and other utilities need to redouble their commitment to eliminating typographical errors and develop more sophisticated algorithms to detect and eliminate the errors. Vendors of integrated library systems also need to develop similar algorithms and spell-check functionality in online library catalogs. Third, utilities need to increase the incentives for enhancing master records by correcting typos,

Table 2. Number of the five sampled libraries with corrected records

Word frequency category									
Very high		High		Moderate		Low		Very low	
Word	#	Word	#	Word	#	Word	#	Word	#
accomodation	2	asessing	2	Adddison	4	0'Donnell*	4	0ccupied*	5
activites	1	Bismark*	1	artifical	1	appendox	1	5oth*	2
amd	3	Carribean	3	bizzare	1	batle	1	autobiograaphy	5
artic	3	charaters	3	Buddist	4	choregraphy	4	Behaviroal	1
cby*	0	Cincinnat	0	commitee	1	comentaries	1	Berkley Calif.	4
Cincinatti	1	classsification	1	Disabilites	2	Comission	2	chlidren	3
commision	5	commom	2	estabishment	2	estalished	2	colleages*	5
commmunity	2	decisons	4	goup	3	inclduing	3	consolidaton	5
environmental	3	Engineeering	2	Havard	3	Januaury	3	dstrict	2
John Hopkins*	2	l865*	0	incoporation	0	nutritution	0	edittion	5
ℓ895*	1	literatue	2	Libray	2	occupatonal	2	Hnery	3
ℓ970*	2	Natonal	0	Miltary	0	peroidical	0	jewlery	4
managment	4	Pennyslvania	3	ocupation	4	pesented	4	microcopes	4
Weidenfeld & Nicholson*	0	peotry	3	Mrs. Polifax*	2	Pesonnel	2	microcopmuters	3
reseach	0	Phillipines	0	proceedngs	1	shinning	1	muusic	4
Tuscon	1	pictoral	1	prodcuts	2	Sulllivan	2	personalites	5
Univeristy	2	poeples	1	rsources	1	surban	1	Pictorialworks	3
Wasington	2	pschological	4	Rusian	2	toliet	2	reseasrch	3
x History*	1	responsiblity	5	Spainish	2	undergradutes	2	VBiography*	1
z United*	4	Russsian	3	supplment	4	wiht	4	worongs	5

* See appendix for correct spellings

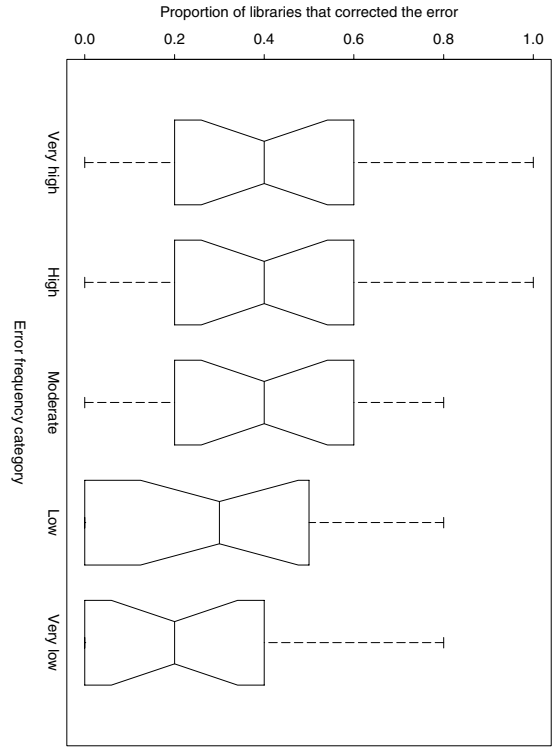

Figure 1. Box-and-whiskers plot of the proportion of errors corrected, by category of error frequency. The first three categories (very high, high, and moderate) show relatively consistent proportions of corrected words, while the last two categories (low, very low) show lower proportions of corrected words.

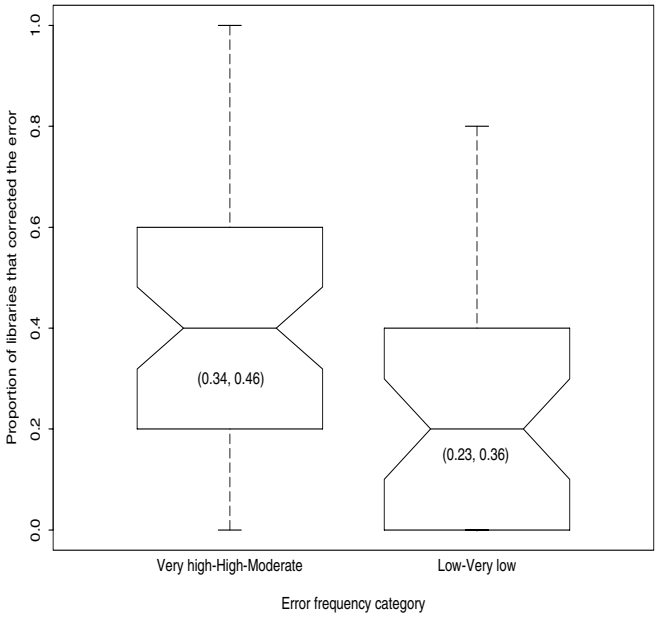

Figure 2. Same as Figure 1, but with the very high, high, and moderate categories combined into one category (sixty words), and the two categories (low, very low) combined into one category (forty words). The 95 percent confidence interval for the true mean proportion corrected for the very high-high-moderate words is (0.34, 0.46). The 95 percent confidence interval for the true mean proportion corrected for the low-very low words is (0.23, 0.36). The 95 percent confidence interval for the difference in these two proportions is (0.02, 0.19), indicating that the observed difference is statistically significantly different from zero with confidence coefficient 0.95.

and they should make it easier for libraries either to correct the typos or to report them to the utilities' quality control departments.

Other areas in the field of bibliographic record error analysis also need to be studied. One valuable direction for future research would be to develop a standard measure of bibliographic database quality. This measure would be based on overall record quality and fullness; number and types of errors present in the records, including not only typographical errors, but also cataloging and authority errors; and size of the database. The resulting database quality rating would aid libraries in planning their database quality control and cleanup work. Perhaps it also would serve as an incentive for libraries to eliminate dirty data from their catalogs and prevent new dirty data from entering them.

Future research also might compare the rate of typographical errors to other types of errors found in bibliographic records, such as errors in subject analysis, errors in the application of the Anglo-American Cataloguing Rules and the Library of Congress Rule Interpretations, and errors in choice of heading. We hope the ability to search many local libraries' online catalogs through the Internet will encourage studies modeled after this one and provide an accurate look at bibliographic data quality in online catalogs overall.

References

1. Sheila Intner, "Quality in Bibliographic Databases: An Analysis of Member-Contributed Cataloging in OCLC and RLIN," *Advances in Library Administration and Organization* 8 (1989): 1–24.
2. Ibid., 12–13.
3. "The Dirty Database Test," *American Libraries* 22 (Mar. 1991): 97.
4. Jim Dwyer, "The Catalogers' 'Invisible College' at Work: The Case of the Dirty Database Test," *Cataloging & Classification Quarterly* 14, no. 1 (1991): 75–82.
5. Ibid., 80.

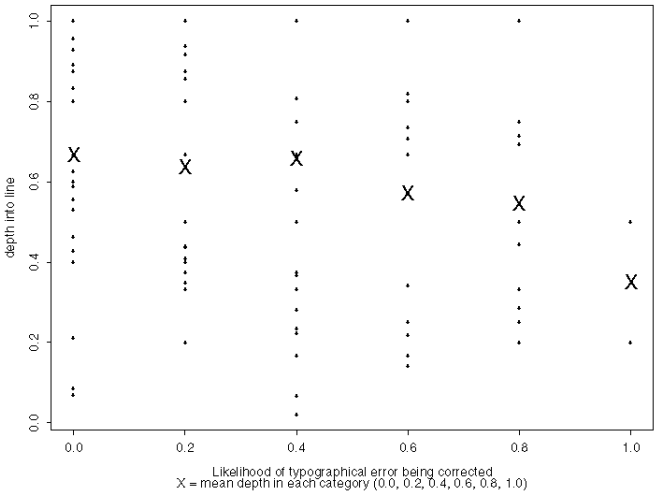

Figure 3. Plot of association between location of word in field ("depth" = (position of word in field)/(number of words in field)) and "likelihood" that typographical error in the word was corrected (0.0, 0.2, 0.4, 0.6, 0.8, 1.0). The capital X's denote the means of the depths for all words in each of the six "likelihood" categories. The lack of any apparent trend in the X's suggests little association between the location of the typographical error in the line and its likelihood of being corrected.

6. Terry Ballard, "Spelling and Typographical Errors in Library Databases," *Computers in Libraries* 12, no. 6 (1992): 14–17.
7. Terry Ballard and Arthur Lifshin, "Prediction of OPAC Spelling Errors through a Keyword Inventory," *Information Technology and Libraries* 11 (June 1992): 139–45.
8. Ibid., 142.
9. Ibid.
10. Ibid.
11. Sylvia A. Gardner, "Spelling Errors in Online Databases: What the Technical Communicator Should Know," *Technical Communication* 39 (1992): 50–53.
12. Ibid., 50.
13. Ibid., 52.
14. Ralph Nielsen and Jan M. Pyle, "Lost Articles: Filing Problems with Initial Articles in Databases," *Library Resources & Technical Services* 39, no. 3 (1995): 291.
15. Ibid., 292.
16. Barbara Nichols Randall, "Spelling Errors in the Database: Shadow or Substance?" *Library Resources & Technical Services* 43, no. 3 (1999): 168.
17. David Bade, *The Creation and Persistence of Misinformation in Shared Library Catalogs: Language and Subject Knowledge in a Technological Era* (Champaign, Ill.: Graduate School of Library and Information Science, University of Illinois at Urbana–Champaign, 2002).
18. Ibid., 4.
19. Ibid., 26–27.
20. Ibid., 27.
21. Ibid.
22. Ibid., 5.
23. Ibid., 3.
24. Terry Ballard, *Typographical Errors in Library Databases*. Rev. Mar. 7, 2002. Accessed May 1, 2002, http://faculty.quinnipiac.edu/libraries/tballard/typoscomplete.html.
25. Ibid.
26. W. Nelson Francis and Henry Kucera, *Frequency Analysis of English Usage: Lexicon and Grammar* (Boston: Houghton Mifflin, 1982).
27. Ballard, "Typographical Errors."
28. John W. Tukey, *Exploratory Data Analysis* (Reading, Mass.: Addison-Wesley, 1977), 39–41.
29. George W. Snedecor and William G. Cochran, *Statistical Methods* (Ames, Iowa: Iowa State University Press, 1980).
30. Robert McGill, John W. Tukey, and Wayne A. Larsen, "Variations of Box Plots," *The American Statistician* 32, no. 1 (Feb. 1978): 12–16.

Appendix
List of Words Containing Typographical Errors and the Number Corrected/Not Corrected
(Sorted by frequency, then by word)

Word	Frequency	Tag	Corrected	Not corrected
accomodation	1-very high	650	2	3
activites	1-very high	245	1	4
amd	1-very high	245	3	2
artic	1-very high	245	3	2
cby[1]	1-very high	245	0	5
Cincinatti	1-very high	245	1	4
commision	1-very high	710	5	0
commmunity	1-very high	610	2	3
enviromental	1-very high	650	3	2
John Hopkins[2]	1-very high	245	2	3
l895[3]	1-very high	100	1	4
l970[3]	1-very high	245	2	3
managment	1-very high	240	4	1
Weidenfeld & Nicholson[4]	1-very high	260	0	5
reseach	1-very high	245	0	5
Tuscon	1-very high	111	1	4
univeristy	1-very high	711	2	3
Wasington	1-very high	650	2	3
x history[5]	1-very high	651	1	4
z United[5]	1-very high	650	4	1
asessing	2-high	245	2	3
Bismark[6]	2-high	651	1	4
Carribean	2-high	650	3	2
charaters	2-high	600	3	2
Cincinnat	2-high	245	0	5
classsification	2-high	650	1	4
commom	2-high	650	2	3
decisons	2-high	245	4	1
Engineeering	2-high	710	2	3
l865[3]	2-high	245	0	5
literatue	2-high	650	2	3
Natonal	2-high	651	0	5
Pennyslvania	2-high	245	3	2
peotry	2-high	245	3	2
Phillipines	2-high	650	0	5
pictoral	2-high	610	1	4
poeples	2-high	245	1	4
pschological	2-high	245	4	1
responsiblity	2-high	710	5	0
Russsian	2-high	650	3	2
Adddison	3-moderate	700	4	1
artifical	3-moderate	650	1	4
bizzare	3-moderate	245	1	4
Buddist	3-moderate	650	4	1
commitee	3-moderate	710	1	4
Disabilites	3-moderate	710	2	3
estabishment	3-moderate	245	2	3
goup	3-moderate	520	3	2
Havard	3-moderate	710	3	2
incoporation	3-moderate	245	0	5
Libray	3-moderate	710	2	3
Miltary	3-moderate	651	0	5
ocupation	3-moderate	651	4	1
Mrs. Polifax[7]	3-moderate	245	2	3
proceedngs	3-moderate	245	1	4
prodcuts	3-moderate	650	2	3
rsources	3-moderate	710	1	4
Rusian	3-moderate	245	2	3

Appendix (continued)

Word	Frequency	Tag	Corrected	Not corrected
Spainish	3-moderate	650	2	3
supplment	3-moderate	245	4	1
0'Donnell[8]	4-low	100	4	1
appendox	4-low	245	2	3
batle	4-low	650	4	1
choregraphy	4-low	246	2	3
comentaries	4-low	245	0	5
Comission	4-low	710	2	3
estalished	4-low	500	2	3
inclduing	4-low	245	0	5
Januaury 4-low	245	0	5	
nutritution	4-low	246	0	5
occupatonal	4-low	245	3	2
peroidical	4-low	245	0	5
pesented	4-low	245	1	4
Pesonnel	4-low	110	3	2
shinning	4-low	700	0	5
Sulllivan	4-low	600	4	1
surban	4-low	245	1	4
toliet	4-low	500	0	5
undergradutes	4-low	245	1	4
wiht	4-low	245	2	3
0ccupied[8]	5-very low	245	0	5
5oth[9]	5-very low	245	3	2
autobiograaphy	5-very low	245	0	5
Behaviroal	5-very low	710	4	1
Berkley Calif.	5-very low	111	1	4
chlidren	5-very low	246	2	3
colleages[10]	5-very low	245	0	5
consolidaton	5-very low	245	0	5
dstrict	5-very low	650	3	2
edittion	5-very low	500	0	5
Hnery	5-very low	245	2	3
jewlery	5-very low	650	1	4
microcopes	5-very low	245	1	4
microcopmuters	5-very low	650	2	3
muusic	5-very low	650	1	4
personalites	5-very low	740	0	5
Pictorialworks	5-very low	650	2	3
reseasrch	5-very low	500	2	3
vBiography[5]	5-very low	650	4	1
worongs	5-very low	245	0	5

1. The correct form is: |c by. The subfield delimiter was left out, and the letter "c" was attached to the word "by."
2. The correct form is Johns Hopkins.
3. The letter "l" was input instead of the numeral for one.
4. This is the name of a publisher. The correct form is Weidenfeld & Nicolson.
5. These are tagging errors. The subfield delimiter "|" was left out.
6. The correct spelling is Bismarck, as in Bismarck, North Dakota.
7. The correct spelling is Mrs. Pollifax. This is the name of a fictitious character.
8. The first letter in the word was entered as a zero (0) rather than an uppercase "O."
9. The zero (0) in 50th was mistakenly entered with a lower-case "o."
10. The correct spelling is colleagues.