Notes on Operations

Spelling Errors in the Database: Shadow or Substance?

Barbara Nichols Randall

The purpose of this research was to determine the extent of spelling errors in the University at Albany's online catalog, whether these errors seriously affect users' access to library materials, and what effect spelling errors will have on the group database planned for the State University of New York. Using standard database tests, I studied the catalogs of the four University Centers (Albany, Binghamton, Buffalo, and Stony Brook) as well as two comparison catalogs: the New York State Library's Excelsior and the University of California's Melvyl. The results of these studies show that misspellings are unavoidable due to the way that most catalogs were built. These errors, however, are rarely an impediment to retrieval. I conclude with suggested ways to find and correct misspellings without expensive large-scale efforts.

> A typographical error in a ship mortgage prepared by Haight, Gardner, Poor & Havens could cost the Prudential Insurance Co. of America between $11 million and $31.5 million before a dispute in federal court is finally resolved . . . at issue is a $92.8 million lien . . . three zeroes were dropped from the amount when the mortgage was amended in April 1986, leaving Prudential with a lien that may be worth only $92,885.—Frost and Goldner (1988, 7).
>
> Damn construction 92 percent complete at Brushy Creek—Leno (1998).

Misspellings and typographical errors in library databases are neither as costly as those made in financial documents nor as funny as those highlighted every Monday night by Jay Leno on *The Tonight Show*. A literature review shows that misspellings and typographical errors have been, however, the subject of much research. Bourne (1977), Ryans (1978), Dwyer (1991), Ballard and Lifshin (1992), Gardner (1992), and Cahn (1994) all deal with the identification or effect of misspelling on a database. Bourne (1977) concentrated on misspellings of index terms, including the number of misspelled terms in computer databases, the implications of these misspellings to searchers, and who should clean up the errors. He found the occurrence of misspelled terms ranging from 1 in 8,000 citations in one database to 1 in 160 citations in another.

Ryans (1978) studied the accuracy of 700 records in the OCLC Online Computer Library Corporation, Inc., database

BARBARA NICHOLS RANDALL is assistant director of Guilderland Public Library, Guilderland, NY.
Manuscript received December 12, 1998; accepted for publication March 1, 1999.

using *Anglo-American Cataloguing Rules (AACR)*, International Standard Bibliographic Description (Monographs) (ISBD (M)), and OCLC input standards as measures. Most of the errors she found were "due to simple carelessness" (131). She found errors on 283 records, including misspellings and typing errors. She did not quantify the errors, but she did describe them as "frequent."

At the time of these early studies, online catalogs were a dream rather than a reality; libraries throughout the country joined shared cataloging networks and began the preliminary work for the eventual computerization of their catalogs. In New York state, a number of large research libraries with adequate money and institutional computer expertise created their own online catalogs. Throughout the 1970s and 1980s, librarians performed extensive retrospective conversion of their card catalogs (Reed-Scott 1985). Federal, state, and local funds were used throughout the country, and catalog records were created by contractors and in-house catalogers, and through national cooperative projects such as the COMARC (Cooperative MAchine Readable Cataloging) and the CONSER (CONversion of SERials) projects. Catalogs of varying quality often resulted, which led to projects to clean up the data.

Beall (AL Aside 1991) started a dialogue on misspelling that is ongoing, with almost quarterly discussions occurring on AUTOCAT. Beall searched the occurrence of 10 common word misspellings, totaled the occurrence of the words, eliminated the *i.e.* or *sic* words, subtracted the total from 100, and compared the result to other libraries of similar size. The 10 misspellings are: Febuary, Guatamala, Misssion, Goverment, Fransisco, Grammer, Recieve, Wensday, Seperate, Conditons. Dwyer (1991) further refined the method by deriving a way to measure a meaningful error rate by comparing the number of misspellings to the number of correct spellings of the words. Cahn (1994) used a measure to take into account whether access was prevented because of the uniqueness of the error or whether the error was redundant and therefore did not affect access to the record. Ballard (1992) published a list of commonly misspelled words in online catalogs as a result of a project begun in 1991 to rid the Adelphi University database of obvious typographical errors.

What does all this discussion of misspelling and typographical errors mean? Does the discussion represent merely the perspective of good spellers and proofreaders? Are our catalogs so flawed that our patrons won't find what they need? According to psychologist Craig Brod (1984, 15), "Unwittingly, we are adopting as our own the computer's standards. We have come to expect from people the perfection, accuracy, and speed to which computers have made us accustomed." As librarians, we must attempt to separate the substance—errors that deny access to information—from the shadow—machine-like perfection.

Purpose

This study was conducted to determine, first, how dirty the University at Albany's catalog data are, and second, the effect that these data will have on a group database planned for the State University of New York libraries (SUNYConnect). "Dirty" was defined here to have two meanings: first, the number of misspellings that occur in the database; and second, the degree to which misspellings inhibit access to the library's materials.

Method

First, the research team searched the University at Albany's catalog using the list of words from Beall (AL Aside 1991). We performed keyword searches and computed a Beall score to get the frequency of error. We next calculated the error rate defined by Dwyer (1991). Then we compared both values to the values found at the two other university libraries in our proposed group database that have similar-sized collections (based on self-reported data from the 1998–99 *American Library Directory*)—Binghamton and Stony Brook. After this preliminary comparison, we searched a set of words in one subject area (economics) both to locate misspellings and to determine

TABLE 1
FREQUENCY OF ERRORS FOR TERMS IN BEALL'S LIST

	Albany	Binghamton	Stony Brook
Febuary	2	4	3
Guatamala	5	0	3
Misssion	0	0	0
Goverment	9	4	16
Fransisco	4	4	16
Grammer	11	7	11
Recieve	1	1	0
Wensday	0	0	0
Seperate	22	23	11
Conditons	6	3	1
TOTAL	60	45	51

the relative importance of the errors found. Finally, the error rates at the fourth library in the proposed group were compared to the rates at the New York State Library, which has a collection of a similar size.

RESULTS

Table 1 presents the Beall scores for the three institutions. Table 2 presents the Dwyer scores for these same institutions. None of the numbers seem to be consequential, but because an important part of database maintenance is correcting errors, the question can be raised about when such errors might be ignored. The answer lies both in the placement of the error—that is, whether the error denies access or whether it does not—and in the uniqueness of the term in the record. To address this, the second stage of our research involved searching for variations on three words related to one subject: economy, economic, and economics. All misspelled variations of the words were found in the Albany, Binghamton, and Stony Brook databases. We found 16 variations in misspelling (see table 3): Albany had 12 spelling variations,

TABLE 2
DWYER'S RATIOS

	Albany			Binghamton			Stony Brook		
	Errors	Total	Rate	Errors	Total	Rate	Errors	Total	Rate
Febuary	2	10,545	5,272.5	4	7,849	1,962.3	3	2,995	998.3
Gautamala	5	1,005	201	0	695	0	3	804	268
Misssion	0	1,298	0	0	1,590	0	0	916	0
Goverment	9	62,640	6,960	4	59,532	14,883	16	39,486	2,467.9
Fransisco	4	5,078	1,269.5	3	9,679	3,226.3	6	9,444	1,574
Grammer	11	5,189	471.7	7	5,151	735.9	11	4,240	385.5
Recieve	1	145	145	1	153	153	0	17	0
Wensday	0	416	0	0	452	0	0	119	0
Seperate	22	1,964	89.3	23	1,876	81.6	11	875	79.5
Conditons	6	28,488	4,748	3	35,021	11,673.7	1	23,187	23,187

Binghamton had 13, and Stony Brook had 9. The same variations in the same bibliographic records occurred in more than one database in 11 instances.

The 16 misspellings all were typographical errors—that is, errors in transcription, not misspellings in the original version. Typographical errors are not solely a byproduct of the computer age, but in fact have existed since early manuscripts were copied letter by letter in monasteries. Alfred Watts, a "printer's reader" (proofreader), wrote a classic work on typographical errors in 1883. Smith (1985) found Watts's work useful in understanding how to improve data entry and proofreading. Smith noted that Watts evaluated each type of typographical error found in a sample of 60 two-column pages of small type set by six different compositors. Watts classified errors into three categories: errors of omission, substitutions, and doubling. Gardner (1992) further refined the substitution and omission categories by extracting two additional categories: errors of letter transposition and errors of letter insertion. An error of transposition occurs when two adjacent letters are interchanged. Errors of insertion occur when an extra letter, either the same or different, is added to the word. Ballard and Lifshin (1992) identified typographical errors as errors of omission, substitution, insertion, transposition, added space, and dropped space.

Using Watts's error categories, the sample contained 6 instances of omission, 8 of substitution, and 2 of doubling in the sample (see table 4). When we consider Gardner's modifications, two of the substitution errors could be called errors of letter transposition and one of the two doubling errors could be called an insertion error. Finally, we did not conduct tests for Ballard and Lifshin's dropped space or added space errors. Spacing problems can be found in some catalogs using forms of internal truncation. However, we did not pursue this approach due to the uncertain results that would be achieved in Albany's database. It is possible that there are undetected instances of both within the databases. As Smith (1985, 189) said, "Although new technology presents new pitfalls for compositors and proofreaders, the old ones—the ones caused by human imperfection—remain to humble us."

TABLE 3
MISSPELLED TERMS RELATED TO ECONOMICS AND THEIR FREQUENCY

Term	Albany	Binghamton	Stony Brook
ecomomy	0	2	0
ecomomic	6	5	0
ecomonic	7	5	1
econonic	0	0	2
econmic	4	4	6
econimine	1	1	2
ecnomic	2	2	1
econommic	2	1	0
economnic	2	0	0
ecomomics	0	2	4
ecomonics	1	3	4
econonics	1	0	0
econmics	2	4	0
economcs	1	1	2
ecnomics	1	1	1
ecoomics	0	1	0

We identified unique errors as such if the misspelled word occurred only once in the record. Following Cahn's (1994) definition

TABLE 4
TYPOGRAPHICAL ERRORS CATEGORIZED USING WATTS' ERROR TYPES

Omissions	Substitutions	Doublings
Econmic	Ecomomy	Econommic
Ecnomic	Ecomomic	Economnic*
Econmics	Ecomonic**	
Economcs	Econonic	
Ecnomics	Econimic	
Ecoomics	Ecomomics	
	Ecomonics**	
	Econonics	

*also insertion
** also transposition

of redundancy, if the misspelled word appeared correctly spelled in another place in the record we classified it as redundant. Albany's database had 40% unique errors (12 of 30) and the remaining errors were redundant. Binghamton's database had 18% unique errors (9 of 49); while Stony Brook had 43% unique errors (10 of 23) (see table 5).

Unique errors are not all equal. For example, errors in a title or subject field are more serious than errors in a note. In a study of online catalog use for the Council on Library Resources, Larson (1983) concluded that most users search by subject. Anderson (1995) reiterates that users rely on keyword and subject searching to find information. Ballard and Lifshin (1992) found the majority of the errors in their study in title fields (63%), followed by note fields (21%), author errors (9%), and series errors (7%). The errors we found in this study occurred in five field types: author fields (including main author, alternate author, and publisher), title fields (including main and alternate titles), subject fields, note fields, and series fields. By analyzing the unique errors, we found that the majority of errors occurred in note fields, with title fields taking second place. There were no errors in subject fields. Redundant errors also followed this pattern.

As we enter the new century, the importance of regional or virtual catalogs has grown rapidly, which has implications for database errors. After the cleanup of an individual catalog, will the creation of a group catalog bring back the errors or compound the errors of our individual catalogs? We wanted to know the overlap of common misspellings in the Albany, Binghamton, and Stony Brook databases. Six of the misspellings occurred in all three databases; 5 of the 6 misspelled words had common records for more than one institution (see table 6). One record was common to all three databases. Eight records were common in two databases. The overlap of some of the typos was bothersome. We wanted to know whether these errors were all from data entry or whether some were the result of a common record that had errors. To check for this in the databases of the three institutions, we used the subset of economics records that existed in more than one database.

In addition to the common misspelled records, common records where one library corrected the database misspelling also existed. Of the total of 16 misspelled economics words, a total of 85 records had misspellings, and 23 records did not have the misspelling. Of the 17 common misspellings (two or more) in table 5, we found 4 records with the terms spelled correctly.

The next step was to look at the catalog of the fourth university center in the proposed group, Buffalo. We ran the Beall, Dwyer, and misspelled economics terms tests on Buffalo's catalog. The Beall score was -40, and the Dwyer scores were considerably lower. The economics terms test revealed 12 misspellings. Of these, 1 was a misspelling not previously identified; 4 were misspellings also found in one other catalog; 2 were misspellings found in two other catalogs; and the remaining 5 were misspellings found in three other catalogs. We identified 3 additional records as common records and found 1 additional misspelling (ecomomy).

TABLE 5
UNIQUE AND REDUNDANT ERRORS BY LOCATION IN RECORD

Database	Error Type	Author	Title	Subject	Note	Series
Albany	Unique	2	5	0	3	2
	Redundant	3	11	0	2	2
Binghamton	Unique	1	3	0	5	0
	Redundant	2	6	0	32	0
Stony Brook	Unique	2	0	0	5	3
	Redundant	3	2	0	3	5

TABLE 6
OVERLAP OF ERRORS AT ALBANY, BINGHAMTON, AND STONY BROOK

Error	Title	Date	Databases
ecomonic	Economic progress	1955	Albany, Binghamton
econmic	A high-speed passenger rail system for the U.S.	1981	Albany, Binghamton
	The economics of direct employment	1900	Binghamton, Stony Brook
	Interest as a source of personal income and tax revenue	1955	Albany, Stony Brook
econimic	Our emergent civilization	1947	Albany, Binghamton, Stony Brook
ecnomic	Miscellaneous essays and addresses	1904	Albany, Binghamton
ecomonics	Philosophy of economics	1982	Binghamton, Stony Brook
	Applied economic forecasting	1971/1966	Binghamton, Stony Brook
	Teachers as agents of national development	1971	Binghamton, Stony Brook

But we were unclear how to evaluate the scores found for Buffalo's database. We were unsure the effect that the size of the collection might have on the results, and thus the relationship between the Beall score of -40 found for Buffalo, and the scores for the other three institutions in the group. As a comparison, we compared Buffalo's scores to the New York State Library, which has a collection of similar size.

The State Library scored even lower on the Beall test, -60, and consistently lower on the Dwyer test. The economics terms test revealed 25 misspellings (see table 7). In comparison to the New York State Library, the quality of Buffalo's database was good.

Why is there such a difference between the State Library and Buffalo? Because it was possible to determine the source of the State Library's records, but not those of the other institutions, those records were studied in depth. The State Library's automated catalog, Excelsior, is a second-generation database. The original database, CMS (Collection Management System), was first operational in 1978 when the State Library moved its primary collections and base of operations from the Education Building to the Cultural Education Center in the Nelson E. Rockefeller Empire State Plaza, the seat of New York State government. The initial retrospective conversion work was done through a contract with a local nonlibrary contractor, Finserv. The State Library was also an early member of OCLC, so OCLC archive tapes were also used. The State Library is one of the original CONSER participants. Additional retrocon projects were performed throughout the 1980s, including: cataloging the American Periodical Series, a grant-funded project to catalog the Goldsmiths'-Kress Library of Economic Literature, two contracted upgrade projects (one again through Finserv and the other through OCLC machine match), and the purchase and direct loading of the SuDocs (Superintendant of Documents) tapes as a full government documents depository. In descending order, the errors originated with Finserv (20), OCLC (19), SuDocs (13), OCLC upgrade (11), Goldsmiths'-Kress (8), Finserv upgrade (6), CONSER (3), in-house direct input (3), access level document cataloging (DACS) (2), and archival records from the State Archives and Records Administration (SARA) (2).

The majority of the errors occurred during retrospective conversion. This is not surprising given that one of the goals of retrocon is always production. All of the titles in the Finserv, Finserv upgrade, and

OCLC upgrade groups were older material matched or upgraded based on shelflist cards. The Finserv project was a separate, production-oriented project undertaken while the move to the new library building was occurring. The library's cataloging staff were not involved in the project initially. The Finserv upgrade and OCLC upgrade projects did have cataloging staff involvement as well as extensive systems evaluation. The inclusion of CONSER records, which underwent rigorous review, in the error group illustrates the very human nature of spelling errors.

OCLC is the bibliographic utility the State Library uses. We wondered whether the errors from the OCLC and OCLC upgrade records still existed in the OCLC database. Thirteen of the 30 records remain misspelled on the OCLC database. This percentage rate, 43%, is lower than the 51% documented by Ballard and Lifshin (1992). Three were in fields input by State Library staff at the time of production and were never present in the OCLC database.

TABLE 7
ERRORS IN ECONOMICS TERMS

Term	Buffalo	NY State Library
ecomomy	1	2
ecomony	0	4
ecmony	0	1
ecomomic	0	10
ecomonic	0	10
econcomic	0	2
econonic	3	4
econmic	13	13
economoc	0	2
econimic	2	1
ecnomic	4	6
econmonic	0	1
econommic	0	3
econoomic	0	2
economnic	2	2
ecocomic	1	0
economoic	1	1
ecomomics	3	6
ecomonics	4	7
econonics	0	1
econimies	0	1
econmics	1	4
economcs	2	1
ecnomics	0	2
ecoomics	0	1

We searched one last database: the University of California (UC) union catalog, Melvyl. Melvyl is a model for the SUNYConnect project. The Melvyl Union Catalog is part of a statewide computer-based library system created in 1981 by the California Digital Library (formerly known as the Division of Library Automation), in conjunction with UC campuses. It has been available online since the mid-1980s (Crowell 1995) and has been available in Web format since 1997. We wanted to compare the error rates found in this almost twenty-year-old catalog.

We searched the Web version of Melvyl using title keyword, subject keyword, personal and corporate author keyword, and series keyword. The note field is not keyword searchable in Melvyl. Melvyl (with 9,678,014 titles and 14,632,800 holdings as of November 25, 1998) is approximately twice the size of the combined university centers, discounting overlap. We found error frequencies for the Beall and the economics term tests. We could not run the Dwyer test because Melvyl does not allow for complete counting of the correctly spelled words. Any Melvyl search that retrieves more than 10,000 hits is stopped.

We found at least one instance of each misspelling. The primary category of errors was subject fields, followed by title, author, and series. Seven of the nine overlap titles in table 6 were also owned by UC libraries and included the same errors.

Finally, due to the way that data entry production is measured, we broke down typographical errors as error per character input, whether written or typed. Watts (Smith 1985) found 1 error in every 1,750 characters. Chan (1994) notes that almost one hundred years later, in July 1980, researchers at the National Composition Association found an error rate of 1 in 1,000.

The University at Albany's error rate is 1 in 1,946 characters.

Conclusion

While perfection in both humans and databases is a worthwhile goal, the reality is that it is also impossible. Most spelling errors are redundant errors and thus, do not prevent users from finding the needed record. As Cahn (1994, 30) correctly stated, "Issues of time and money cannot be ignored." Bourne (1977, 9–10) called misspellings "internal parasites to the search system," yet he said, "while a relatively large number of index terms are misspelled (compared to conventional printed indexes), and while those errors are very conspicuous, they in fact have relatively little impact of file use for many of the databases." Most librarians probably will agree with Ballard and Lifshin (1992, 139), who pointed out, "It may be widely perceived that spelling errors in OPACs and other large databases are few in number, randomly distributed, and impossible to locate in any systematic fashion. . . . every library that has an OPAC with keyword capability should search the problem words that we have identified and fix the inevitable errors."

In consultation with the head of cataloging at the University Libraries, we chose to take a staged approach to database cleanup. Although the consensus was that the errors were minor, we wanted to search for the terms in Ballard's list. Because we lacked staff time and money, a volunteer conducted the search for us. The volunteer noted only the number of occurrences of the misspelled terms and found 697 potential misspellings of 106 words on the Ballard list. Some misspellings were in fact correct transcriptions of title page errors and on investigation were correctly labeled "i.e." or "[sic]." There were no instances of misspellings for 73 of the words found on the Ballard list. This list is being used as a guide for correction as staff or students become available to do the work. We are correcting the most frequently misspelled words first.

Concern about the impact of misspellings on the catalog should be minor. Although misspellings or typos are embarrassing, the low number of unique occurrences of each misspelled term means that in most cases a user will still be able to find a relevant item. The breakdown of the State Library's misspelled records suggest that most of the misspellings or typos might be traced to the profession's early years of retrospective conversion. Given that this work can be sorted out and listed by project code, it becomes possible to target these records for further examination. Moreover, some of those records will disappear as we inventory and weed our collections.

Reports from our systems people bode well for our quest for perfection. Each time we request a report to work on an area of problems in the catalog, we also find a few other problems, usually misspellings. Recently, when requesting a report of all the unbracketed general media designators in the title transcription area, we also generated a list of instances of "micorform," "midroform," and "videorcording." We are aided in our perfection quest by sharp-eyed spellers who send our department errors they notice in the catalog.

However, correcting spelling errors and typos can take more time than most libraries have. The small number of these errors that can be reduced further (given the concept of uniqueness) shows that the type of large-scale effort Ballard (1992) performed at Adelphi University is beyond the means and needs of most libraries. The combination of errors that occur as union catalogs, whether virtual or otherwise, are created increases the number of errors but not by any consequential amount. We believe that our current error-correction efforts, on an as-needed basis or as a byproduct of other enhancement projects, are both sufficient and reasonable.

Works Cited

AL Aside – Ideas. 1991. *American libraries* 22: 197.

American library directory 1998–99, New Providence, N.J. : R. R. Bowker.

Anderson, Judy. 1995. Have users changed their style? *RQ* 34: 362–68.

Ballard, Terry. 1992. Spelling and typographical errors in library databases: One library's

system for rooting out spelling errors. *Computers in libraries* 12, no. 6: 14–19.

Ballard, Terry and Arthur Lifshin. 1992. Prediction of OPAC spelling errors through a keyword inventory. *Information technology and libraries* 11: 139–45.

Bourne, Charles P. 1977. Frequency and impact of spelling errors in bibliographic data bases. *Information processing & management* 13: 1–12.

Brod, Craig. 1984. *Technostress: The human cost of the computer revolution,* Reading, Mass.: Addison-Wesley.

Cahn, Pamela. 1994. Testing database quality. *Database* 17, no. 1: 23–30.

Crowell, Susan. 1995. The MELVYL online public access catalog: Library and remote usage trends. Available: http://www.wizweb.com/~susan/melvyl.html. Accessed: December 9, 1998.

Dwyer, Jim. 1991. The catalogers' "invisible college" at work: The case of the dirty database test. *Cataloging and classification quarterly* 14, no. 1: 75–82.

Frost, Edward and Diane Goldner. 1988. Typo could cost millions. *Manhattan lawyer,* March 29–April 4, 1988: 7.

Gardner, Sylvia A. 1992. Spelling errors in online databases: What the technical communicator should know. *Technical communication* 39, no. 1: 50–55.

Larson, Ray R. 1983. *Users look at online catalogs, part 2: Interactivity with online catalogs, final report for the Council on Library Resources.* Washington, D.C.: Council on Library Resources.

Leno, Jay. 1998. Headlines. Available: http://www.nbc.com/tonightshow. Accessed: December 9, 1998.

Reed-Scott, Julia. 1985. Retrospective conversion: An update. *American libraries* 16: 694–98.

Ryans, Cynthia C. 1978. A study of errors found in non-MARC cataloging in a machine-assisted system. *Journal of library automation* 11: 125–32.

Smith, Peggy. 1985. Typos yesterday, today, and tomorrow. *Scholarly publishing:* 175–89.