5-1997

OPACs and Typograhpical (Sic) Errors

Terry Ballard

Gill Library, College of New Rochelle, tballard@cnr.edu

Follow this and additional works at: http://digitalcommons.cnr.edu/gill-publications

Part of the Library and Information Science Commons

The Systems Librarian

by Terry Ballard

OPACs and Typograhpical (Sic) Errors

The Internet could benefit from some systematic spell-checking as well

Sometimes big things start out small. Back in 1991, I was working as the systems librarian at Adelphi University. Every so often, my boss, Jacqueline Maxin, associate dean for technical services, would send me a screen dump from our online catalog with a misspelled word circled in red. No problem. I could go into the system's record update mode and correct the word in a minute. Every time I did this, I assumed that the last misspelled word in the OPAC had been corrected. Then I would get another one.

One day I did an experiment. I typed the nonsense word "AAAAA" into the OPAC and got a screen of nearby hits. I pressed F to go forward to the next screen, and found that I could go through as many screens as I wanted, seeing all of the indexed words in the catalog. It took about 90 minutes to get through the letter A, and in the process I found dozens of misspelled words needing correction. Many of them were low-level problems—words in the notes fields. Sometimes they were in the title field, and occasionally the problem word would be the first word of the title. In that case, the record is invisible to a correct title search.

Originally, my assumption was that the typos I found would be random single occurrences. However, before I got through the A's, I found the word "administration" with 10 hits. There were many other multiple hits for bad words, although "adminstration" would turn out to be the worst example in the Adelphi catalog. I showed this to Dean Maxin and got the go-ahead to look through the entire keyword index at the rate of one letter a day and pass the problems along to a student assistant for correction.

By the time I got to C, I happened upon the word "commerical" with eight hits. This turns out to be the bellwether example for testing the purity of any database. When I went to a neighboring university and found more than 40 hits for "commerical," I knew I had information that other libraries could use, and I started compiling the error sheets for eventual

publication. By the time I got to "ZZZZZ" (meaning the end of the alphabet, as opposed to falling asleep at my desk), there were more than 1,000 records of the 250,000 in the catalog that needed to have a word corrected.

Getting the Word Out

A colleague—science librarian Arthur Lifshin—and I wrote the findings up as a scholarly article, analyzing the mistakes right down to the parts of speech. (That was mainly Arthur's doing. To me, anything beyond noun, adjective, verb, adverb, and article is probably a preposition and of no interest to me.) I also volunteered to talk about this at the Computers in Libraries conference in Washington early in 1992.

That was my first talk at the national level, and it was going fine as I told about how I found and corrected these problems in our own OPAC. When I said that most of these records came from bibliographic utilities, so the same problems are found in all OPACs, the response was electric. More than a hundred systems librarians were writing the more important words from my transparencies so they could go back to their workplaces and find out that I was correct. The program moderator, Eric Flower, promised to publish the entire list of words in the June 1992 issue

of *Computers in Libraries* magazine. The article would later win the journal's article of the year award, which was presented at the next year's conference.

So far, it seemed to be a totally win-win situation. I cleaned up our online catalog and disseminated this information so that others in the field could use it. Imagine my surprise when this turned out to be somewhat controversial. When I first began to publicize these findings, a report was printed in the *Chronicle of Higher Education*. Unfortunately, the headline gave some people the impression that my university had a poorly maintained database. This was seized upon by a professor at Adelphi during a faculty meeting, and the next week I had to appear before the faculty senate to explain that I had uncovered a problem of international proportions, going far beyond our own database. That flap blew over fairly quickly, but in the years since, I have heard from a few nonlibrarian computer workers that this is just trivia. "There is always be mistakes. So what?" I am *not* making this up.

Librarians Guard the Word

The problem is that these people are living in the here-and-now, while librarians are working to preserve ideas from times past and send them carefully into

the future. I can't speak for everyone in the field, but when there is a problem in my database that can be corrected, I correct it. So much work goes into the production of a MARC record that it is unthinkable to live with an error in the first word of the title, which makes the whole thing invisible to searchers.

In the years since, I have heard from many librarians who have used my work to keep their own OPACs clean. People have even brought that up when I was at job interviews. Most recently, Pat Ensor at the University of Houston contacted me and requested my list for a book she was editing, *The Cybrarian's Manual*, published by ALA early this year. If you want a quick assessment of how any database is doing, try the following words in addition to the ones I already mentioned: acquistion, associaton, behaviorial, bibiography, Britian, psychanalytic, questionaire, reseach, univeristy, universty.

Even knowing that they are all wrong, you may have had to look twice at some of those words to find the mistake. Imagine how easy it is for them to get by even the most careful proofreader, especially knowing that the same words are usually spelled correctly in the same record repeatedly. The only way to root out these problems is to go after them systematically, either using a list or running the database through a systematic spell checking program.

Meantime, people adding pages to the Internet could benefit from these findings. The last time I looked in HotBot, there were 19,000 hits for "commerical" and a staggering 37,000 hits for "infomation," which seems to be the most misspelled word on the Internet.

However, I was recently encouraged by the care that some vendors take with their data. At last year's National Online Meeting, I talked with representatives from a news service in Japan. I made a keyword search of "commerical" and only got one hit in their entire database of Asian business information. They were absolutely horrified that their service would have even a single mistake. If everyone took that kind of care, it would be a happier world.

Terry Ballard is automation coordinator at the New York University School of Law Library. He can be reached by e-mail at ballardt@turing.law.nyu.edu.

> ## "
> *The last time I looked in HotBot, there were 19,000 hits for 'commerical' and a staggering 37,000 hits for 'infomation,' which seems to be the most misspelled word on the Internet.*
> ## "

EBSCO Publishing Integrates EBSCO*doc* with EBSCO*host*

EBSCO Publishing has announced the integration of the EBSCO*doc* document delivery service with EBSCO*host*. Now, EBSCO*host* customers with EBSCO*doc* accounts can order full-text documents while conducting research within EBSCO*host*.

Customers can request the delivery of articles chosen from more than 31,000

periodical titles and conference proceedings found in the multiple databases of EBSCO*host*. A request automatically generates an order, complete with the customer's EBSCO*doc* account preferences and article information, which is forwarded to EBSCO*doc* for fulfillment.

"The accessibility of EBSCO*doc*'s document delivery service from within

EBSCO*host* complements EBSCO Publishing's offering of full-text databases and ensures complete coverage of scientific, technical, and medical literature," said Kimberly Voltero, product manager.

This document delivery service is available on all platforms of EBSCO*host*, a Z39.50-compliant online periodical reference system, providing intuitive, flexible access to bibliographic and full-text databases via a Web browser, Windows, or character cell client.

EBSCO*doc* is a leading document delivery provider operating an on-site service in the University of California

Berkeley Library and utilizing imaging technology to enable a rapid turnaround of document orders. Orders can be placed via fax, e-mail, or online and delivered via Ariel Internet fax, U.S. mail, or courier. Rush services are also available. Customers may place cost and/or time limits on individual orders or on all incoming orders, and they can also receive management reports on a regular basis.

Source: EBSCO Publishing, Ipswich, MA, 800/653-2726, 508/356-6500; Fax: 508/356-6565; http://www.epnet.com.