Abbreviations, Full Spellings, and Searchers' Preferences

By Jeffrey Beall
Auraria Library, University of Colorado Denver, Denver, Colorado, USA

The publisher's version of this article is located here:
http://www.tandfonline.com/doi/abs/10.1080/01639374.2011.595886

Abstract

This study examined ten, selected word pairs, each containing a word's full spelling and its abbreviation, to determine which form search engine users preferred in searching. Using seven search logs gathered from several internet search engines with approximately 608 MB of data, the study measured the occurrences of the twenty terms. The selected words are important in library cataloging, for some are prescribed abbreviations in metadata content standards. The study found that in eight of the ten word pairs users preferred to search words' full spellings over the abbreviations, often by a high margin.

Keywords: Abbreviations, Cataloging, Content standards, Online searching
The author wishes to thank Xiao Guo for her help in the preparation of the tables.

Introduction

Abbreviations have always been an impediment to successful online searching. When a document refers to a concept only by its abbreviation, and a searcher searches the fully-spelled-out form of the word, there may not be a match, and relevant documents may be excluded from the search results. Conversely, when a searcher searches on the full form of a term and relevant documents only include the shortened form, valuable results can be missed.

Research Question

This study seeks to answer to this research question: Do searchers more frequently search on the fully-spelled-out forms of certain words, or do they more frequently search on the abbreviated forms? Knowing the answer to this question could be helpful in several ways. For full-text search engines,

the answer could help better inform query expansion systems. Query expansion occurs when an information retrieval system searches more than what the searcher enters in the search box. In the case of abbreviations, the search engine could add the full form of a word when a searcher enters an abbreviation. For example, when a searcher enters "Nev" the search engine would add the Boolean expression {OR Nevada} to the search.

Increasingly, metadata databases are searched as if they were full-text databases. Instead of searching individual author, title, or subject indexes, searchers perform keyword searches in metadata databases such as online public access catalogs. Depending on the content standard used to encode the metadata, abbreviations may be present in the metadata database. Knowing searchers' behavior in searching terms that can be abbreviated can help better inform the creators of information standards, including metadata content standards.

Also, as this is the first research project of its type, the author seeks to determine the how future research on this topic might be focused, what pitfalls to avoid, and how to generate results that might be more generalizable.

For simplicity, most of this paper refers to all shortened forms of words, including acronyms, initialisms, etc., as abbreviations.

Abbreviations and Cataloging

Cataloging content standards have long prescribed the use of abbreviations. The standards and cataloging rules originally prescribed abbreviations to save space on catalog cards. During the transition from catalog cards to online catalogs, the content standard, *Anglo-American Cataloguing Rules*, 2nd edition (AACR2) [1], did not change in terms of standards for abbreviations. AACR2's appendix B lists prescribed abbreviations for use in bibliographic records. The abbreviation rules apply to terms catalogers transcribe in the descriptive elements of bibliographic records, such as *ed.* for *edition*. The rules also apply to words used in headings, especially geographical place names. For example, AACR2 prescribes that *Colorado* be abbreviated as *Colo.* when it appears as a qualifier in a corporate name heading. Controlled vocabularies also prescribe the use of abbreviations in headings for subject metadata. For

example, the Library of Congress Subject Headings also uses abbreviations for geographical place names.

One popular example of a word / abbreviation pair is *department* / *dept.* AACR2 does not prescribe abbreviating the word *department* in headings, but the corresponding Library of Congress rule interpretation does [2]. In late August, 2010, the Library of Congress asked for comments on changing existing policy to strike the rule interpretation that prescribes abbreviating *Department* to *Dept.* Later, in January 2011, the Library of Congress, having closed the comment period for this proposal, suddenly announced that it would not make a change in the rule interpretation regarding the abbreviation of *Department*. The announcement stated, "The few comments received by the Policy and Standards Division, Library of Congress via email showed a clear preference for making this change but the limited response did not constitute a mandate" [3].

However, the emerging cataloging content standard, *Resource Description and Access* [4], almost completely does away with prescribing the use of abbreviations in bibliographic records. The list of abbreviations that RDA prescribes are in Appendix B.7 of the code, and the list is smaller than AACR2's and is chiefly for non-English language terms. Some of the few prescribed abbreviations that remain in RDA are *v.* for *volume*, *no.* for *number*, and *in.* for *inches*.

Book and other resource titles often include abbreviations, but generally cataloging codes prescribe that a title added entry be added to the bibliographic record with the spelled-out forms of the terms, especially if they occur in the first few words of the title. This practice insures that a searcher will find the work in an online catalog regardless of the spelling use – full or abbreviated.

The Nature of English Language Abbreviations

Abbreviations are shortened forms of words or phrases. For example, *Dr.* is the abbreviation of *doctor*. Acronyms are words formed by taking the first letter (or letters) of each word in a name or expression and pronounced as a word. *CREATE* is the acronym for the *Center for Research and Evaluation in the Application of Technology to Education*. Initialisms are abbreviations formed by taking the first letter of each word in a name or

expression that is pronounced using the names of the individual letters. *NCAA* is the initialism for *National Collegiate Athletic Association*. Abbreviations, acronyms, and initialisms are synonyms of their counterpart full spellings because they are two terms that mean the same thing and therefore meet the definition of synonym. *DOE* is the initialism for Department of Energy; both terms have exactly the same meaning.

Unfortunately, the shorter a term is, the more likely it is to be ambiguous. In the above example, an information retrieval system might confuse *DOE* with *doe*, a term for a female deer. In this way, problematic synonyms simultaneously become problematic homonyms. *DOE* and doe are homonyms because they are a single spelling of a term that has more than one meaning. Homonyms lower search recall, and synonyms decrease search precision. So a query expansion system that adds either the full spelling or the abbreviation to a search will increase recall but can greatly lower precision.

Other Studies

Few studies have dealt with the problem of abbreviations in information searching, especially in the library science literature. Writing in 2008, I summarized the problem:

Abbreviations, acronyms, and initialisms can hinder recall in full-text search systems because a document may contain only the short form of the word or only the long form. When this occurs, someone searching on the short form (PETA) will miss in his retrieval documents that only use the long form (People for the Ethical Treatment of Animals). Alternately, searching on the long form of the term, like Magnetic Resonance Spectroscopy, will miss documents that only refer to the concept by its short form, MRS [5].

In information science, research on abbreviations generally focuses on resolving the problems they create in information retrieval using algorithmic processes. A recent article entitled "Automatic expansion of abbreviations by using context and character information" [6] is an example. Researchers use the terms "abbreviation expansion" or "abbreviation resolution" to describe the process. Many methods rely on specially-created abbreviation dictionaries that add all the full forms of a document's abbreviations to the existing short forms. This way, if an end-user searches the full or abbreviated form of a term, the system finds a match.

Methodology

The purpose of this study is to find out whether searchers more often search the full or abbreviated forms of a selected list of words. To do this, I decided to search ten full spelling / abbreviation word pairs in search logs and record and analyze the results. I began by examining the search logs generated by my library's Innovative Interfaces, Inc. online catalog. One of the features of the staff mode allows a librarian with authorization to access a record of the catalog's search logs for a period of one week, ending at the time of the download. Typically, search logs contain the text the user entered in the search, along with the index the user searched, such as the author, index, keyword index, etc.

After examining these files, I determined that they were too small. Many of the words I selected to study did not occur in the search logs at all, and others occurred only a few times, a situation that would lead to meaningless search results. So I sought a larger search log. Inquires on email lists and searches in the literature led me to Bernard J. (Jim) Jansen, an associate professor of information sciences and technology at The Pennsylvania State University. He shared with me seven search log files from internet search engines that he has used in his own research. The files were created from searches executed in the Excite, AlltheWeb, and AltaVista search engines during the period 1997-2002. Table 1 lists the files along with their sizes.

[Insert table 1 about here]

Search engines do not generally make their search logs available to researchers. However, Jim Jansen was able to collect these seven files and generally shares them with other researchers. Describing how he acquired one of the files, Jansen wrote,

A panel session at the 1997 ACM Special Interest Group on Research Issues In Information Retrieval conference entitled "Real Life Information Retrieval: Commercial Search Engines" included representatives from several Internet search services. Doug Cutting represented Excite, one of the major services. Graciously, he offered to make available a set of user queries as submitted to his service for research [7].

Excite and AltaVista still exist as internet search engines, but AlltheWeb ceased operations in April, 2011 [8]. Other studies have used the same search logs this study used. For example, Jansen and Amanda Spink used the two AlltheWeb files listed in Table 1 in their article, "An analysis of Web searching by European AlltheWeb.com users" [9].

Jansen is a prolific author of studies that measure transaction log data. He also has written on the methodology of search log analysis [10]. Transaction logs are broader than search logs. Transaction logs, the broader term, may include user navigation logs within a particular website and other searches, but search log analysis is specific to user searching behavior, according to Jansen. He defines search log analysis as "the use of data collected in a search log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes" [11].

The Ten Full Spelling / Abbreviation Word Pairs

Instead of randomly selecting the word pairs to use in the study, I selected word pairs that I thought it would be useful to study, especially in the context of cataloging. I wanted to use some words that catalogers commonly use in bibliographic records. By choosing such words, I hoped to get an idea of the etiology of some of the unsuccessful searches on metadata databases due to the use of abbreviations in the search terms or in the metadata records. Another reason I didn't choose the words randomly is because I was afraid that a random selection would lead to words or terms that are rarely used or that are highly ambiguous. For example, if an abbreviation such as *co* was among the randomly-selected words, the results wouldn't say much because many different words share that abbreviation. I wanted to avoid searching occurrences of ambiguous abbreviations because the results would not be as meaningful as they would be with less ambiguous pairs.

But because I didn't select the words randomly, the results really only apply to the word pairs themselves and are not necessarily generalizable. That is to say, one cannot make a general statement about the proportion of searches using abbreviated or fully spelled-out forms of words based on the results of this study. It is possible that the result is different in every case and unique to each word pair. If this is the case, then selecting rather

than randomly selecting words is an equally valid methodology for studying user preferences. Table 2 lists the word pairs I included in the study.

[Insert Table 2 about Here]

Selection and Description of the Word Pairs

Corporation / Corp. I selected this pair because the full spelling and the abbreviation are quite common. I wanted to include common words in the study. Later I realized that **Corp.** is also the abbreviation for corporal, so the data I gathered for this pair will be helpful in analyzing a pair that includes an ambiguous abbreviation.

Government / Govt. This word and its abbreviation are commonly found in bibliographic records, and cataloging codes often prescribe abbreviating *government*. The word is frequently used in imprint statements and in corporate body headings. I wanted to find out which term searchers prefer.

Limited / Ltd. This word is often used in name headings. In the context of a business, limited is similar to *incorporated.* But the word itself can also be the past tense of *to limit*, so the word itself is a homonym, and the abbreviation is less ambiguous, but the term *LTD* can sometimes refer to an automobile model.

Miscellaneous / misc. I selected this word pair because I wanted a less common word with a reasonably well-known and well used abbreviation. This is an example of an abbreviation, like *Corp.*, that uses letter-for-letter the first few letters of the fully-spelled-out word, unlike abbreviations like *Govt*., which use non-sequential letters to form the abbreviation.

Paperback / pbk. I selected this pair of terms because of its importance to cataloging and bibliographic description. Currently, catalogers often use the abbreviation *pbk.* in bibliographic records to qualify an ISBN.

California / Calif. Again, here is an example of a word / abbreviation pair that has two special characteristics. First, the abbreviation matches the first few letters (in this case five) of the full word, and second, there is more than one abbreviation for the full term. The other abbreviation for *California* is *CA* or *Ca.* There even exists a three-letter abbreviation for the state: *Cal.* It's likely that there were many searches in the search logs that I did not

record because searchers used these alternative abbreviations. So the results for this word pair will only be for these two forms. Searching for *CA* or *Cal.* would be especially meaningless because the terms *CA* and *Cal.* can mean many things other than *California*.

Department / Dept. and **Departments / Depts.** I selected *Department* as one of the words because of its importance in corporate name headings in bibliographic description and because of the recent discussions of this term and its abbreviation in the cataloging community, as described earlier. I thought it would also be valuable to examine the plural of this term and its abbreviation. In both cases, the abbreviations differ from the first few letters of the full word; they are not an exactly shortened form of the word.

Boulevard / Blvd. I wanted additional examples of pairs such as *government / govt*. and decided to select this pair. Here again is an example of an abbreviation that is formed not by taking the first few letters of the word but by taking selected letters from throughout the word. I also selected this pair because I wanted additional common word pairs in the study. Boulevard comes to English from the French, and interestingly, the abbreviations in French (*bvd* and *bd*) differ from the English one. So it's possible that in this case results for the full form will measure results in English and French but results for the French abbreviations will be missed.

Internal Revenue Service / IRS This pair is an example of a multi-word name and its initialism. In this study, I searched both *IRS* and *I.R.S.* as the abbreviations. The abbreviation is highly ambiguous, for there are nineteen corporate bodies in the Library of Congress authority file that have *IRS* as a cross reference.

Gathering the Data

Because of the large size of the search log files, only one application on my computer was able to open them without crashing or exceeding the maximum file size limit: Notepad. Using Notepad, I counted the occurrences of the study words and abbreviations manually. While counting, I needed to make sure the count was complete (no terms missed) and accurate (no false hits counted). For example, when I was measuring the occurrences of the abbreviation *misc.* in the seven search log files, I made sure that I only counted occurrences that were the actual

abbreviation itself. I needed to ensure that fuller spellings of words that included the string *misc* were not counted, words such as the full spelling *miscellaneous* and other words that by chance contain the same string, such as *miscarriage*.

For the abbreviations, I searched the terms both with and without ending punctuation, such as the period or full stop. Generally, the search logs include both the user's search terms and the URL of the website the user selected after examining the search results. I did not count any of the data in the search log URLs. Typically, search logs list the URL the user clicked on after examining the search results. I did not analyze each occurrence to ensure that it was really a valid use of the word or abbreviation. Thus, some misspellings and possibly some foreign language terms are included in the results. For example, I observed that the term *Marine Corps* sometimes appears incorrectly as *Marine Corp*. *Corp.* was one of the abbreviations this study used, and it was counted even when it was a misspelling. Also, the study assumed that each search used only a single form – full spelling or abbreviation – though in practice that may not have been the case.

Results

Table 3 lists the number of occurrences for each word and abbreviation covered in the study. It also lists the totals for each term across the seven search logs (the column called Total 1) and the total occurrences of both the full spelling and the abbreviation together (the column called Total 2).

[Insert Table 3 about here]

Discussion

In all but two cases, the end users searched for the full spelling of the word more often than the abbreviation. In study words for which AACR2 and LCRI prescribe abbreviation, such as *Government*, *California*, and *Department*, searchers preferred using the fully-spelled out forms of the words by wide margins. The word-pairs include words with high occurrences, such as *California / Calif.* (23.449 total searches) and pairs with few occurrences, such as *Miscellaneous / Misc.* which appeared in searches only 113 times.

Discussion of the Results from Each Word Pair [11]

Corporation / Corp.

Together, these two terms occurred 8,110 times in the study, with the full spelling occurring 57.47% of the time and the abbreviation 42.53% of the time. As mentioned earlier, *Corp.* is an ambiguous abbreviation; it can mean other things, such as *corporal*. So it is likely that at least some of the occurrences of *Corp*. did not occur as an abbreviation of*corporation*.

Government / Govt.

I found that searchers prefer to search the word using its fully-spelled out form 98.20% of the time and its abbreviation only 1.80% of the time. Among the word pairs I studied, this the second-most lopsided result, after *Departments / Depts.* The abbreviation Govt. likely appears in many millions of bibliographic records, so this finding is significant because catalogers have been recording a term that doesn't match the way searchers seek it, at least in search engines. A keyword search in a metadata database such as an online catalog that has records containing the abbreviation *Govt.* may be contributing to failed searches. When the term only appears as an abbreviation and searchers search using the fully-spelled out form, a match will not occur, and relevant records will not appear in the search results.

Limited / Ltd.

This was one of the two word pairs in which the abbreviation occurred more frequently than the full spelling, 55.45% to 44.55 percent. Both words have more than one meaning. *Limited* can be the past tense of the verb *to limit*, can follow a company's name, and can have a meaning similar to *incorporated*. The abbreviation *Ltd.* is the simple abbreviation of limited in the incorporated sense and among other things is an automobile model: *LTD*. Given the different meanings, it is difficult to draw a conclusion on this pair. We often observe business signs that almost always say *Ltd.* and not *limited* following a company's name, so one might conclude that for this sense of the terms, the abbreviation also occurs more often in searches.

Miscellaneous / misc.

Together, this word pair occurred only 113 times in the search logs. The full spelling made up 77.88% of the occurrences and the abbreviation 22.12%. The chief use of this term is as an adjective and that perhaps accounts for the low number of occurrences. I don't think that I can make any significant conclusion from this term's results, except that the full spelling occurred more frequently.

Paperback / pbk.

Similar to *miscellaneous* and its abbreviation, this word pair occurred relatively infrequently in the search logs I studied. The total number of occurrences was a mere 127, with the full spelling appearing 88.19% and the abbreviation 11.81% of the time. In MARC records, the abbreviation *pbk.* occurs frequently, but within this study's search logs, it occurred only 15 times.

California / Calif.

The most frequently occurring word pair in the study, *California* in full and abbreviated form occurred 23,449 times. Significantly, the full spelling occurred 98.23% and the abbreviation 1.68% of the time. Given that there is a second highly ambiguous abbreviation that I did not include in the study, *Ca.*, the results here show that the abbreviation *Calif.* is hardly used. Perhaps some searchers, after getting imprecise results using *Ca.* did a second search using the full abbreviation.

Department / Dept. and **Departments / Depts.**

The full spelling of the singular *Department* occurred 87.60% of the time, and its abbreviation 12.40%. This finding is significant for cataloging, where *Dept.* is a prescribed abbreviation according to AACR2. This finding suggests that the term should appear in its full form in bibliographic record, for that is how searchers most commonly spell it. The plural forms of the word, *Departments* and *Depts.*, occurred only rarely. In fact, the word *Depts.* occurred only five times among all seven search logs. At 98.78% of all occurrences of this word pair, the full spelling, *Departments*, is the one that searchers by far prefer.

Boulevard / Blvd.

Together, these words occurred 300 times: 172 times for Boulevard and 128 times for Blvd. In percentages, the full form made up 57.33% and the abbreviated form 42.67% of all occurrences. Searchers used both terms about half of the time. These results highlight the importance of query expansion in full text searching. When a searcher enters the term *Boulevard*, for example, a query-expansion enabled information retrieval system might add a background "(OR blvd)" to the search, to help ensure the search results contain all occurrences of the word, regardless of spelling. In other words, the search process might occur this way:

Searcher enters:	Sunset Boulevard
Search engine searches:	Sunset (Boulevard OR Blvd.)

This is an example of query expansion that will likely increase recall without a significant loss in precision, for both terms have few if any homonyms.

Internal Revenue Service / IRS

This last pair is the second of two word pairs in which the abbreviation occurred more frequently than the fully-spelled out form. In this case, the abbreviation is an initialism that occurs both with and without periods. The number of occurrences for the abbreviation reflects the sum of both spellings: IRS and I.R.S. The full spelling occurred 29.74% of the time, and the shorter form occurred 70.26%. A full-text search against a metadata database that contains bibliographic records with the full form of the name and not the shortened form would have low recall due to the prescribed full spelling in the headings.

Further research

Until information scientists find a way to achieve automated abbreviation resolution in information searching systems, abbreviations will continue to be a hurdle in information retrieval. There are several areas in which research on abbreviations may both help to achieve reliable abbreviation expansion and to help improve existing systems and standards.

First, it would be helpful to carry out research that enables information and library science to make general conclusions about the frequency of abbreviations versus their fully-spelled counterparts in both full text documents and in search queries. This might be done by selecting

abbreviated words randomly from a list and designing an experiment that finds generalizable statements about abbreviations in information seeking. The research ought to determine how serious the problem of abbreviations in information retrieval really is.

Second, this same research ought to be completed on a large volume of library online public access catalog search logs. Perhaps a large university or library consortium could make these logs available for research, decoupling the queries from any patron-identifying data.

Third, abbreviation research provides a good opportunity to apply research directly to metadata content standards. Research findings could be applied directly to these standards, for example, limiting the use of abbreviations, or preferring specific abbreviations that users enter more frequently than their counterpart term. The research may also be applicable elsewhere. For example, dictionary entries are loaded with abbreviations, a practice designed to save space in printing, yet the need to save space becomes obsolete with online publishing.

Conclusion

For centuries, humans have used abbreviations to save space in printed works, including catalog cards. Now that most publishing is done electronically, the need to save space on a printed page is gone. Style guides and metadata content standards ought to consider discontinuing the use of abbreviations, except perhaps when a given abbreviation is more common than its counterpart full spelling. Because acronyms are a useful way to remember and state the name of long organizations, they have value and will continue.

Among the ten full-spelling / abbreviation word pairs we searched, in only two cases did the abbreviated form occur more frequently than the full form. Some abbreviations appeared relatively few times overall, especially given the size of the corpus of search logs the study used. For example, the abbreviation *Misc.* appeared only 25 times, the abbreviation *pbk.* 15 times, and the abbreviation *depts.* five times.

Because these terms we included in the study were not selected randomly, the results apply only to the given word pairs and cannot be generalized. In terms of the presence of abbreviations in online library catalogs, this study

found that searchers most often prefer to use full spellings in searching. This preference is problematic because library catalogs are filled with prescribed abbreviations. Additional research on randomly-selected abbreviations would be a valuable method of recording generalizable results regarding users' preferences in searching words and their counterpart abbreviations. These results will then better inform metadata content standard creators, information scientists involved in query expansion, and vendors and designers of information retrieval systems.

Endnotes

1. Joint Steering Committee for Revision of AACR., and American Library Association, *Anglo-American cataloguing rules*. 2nd ed. (Washington: Library of Congress, 2002).

2. Library of Congress, *Library of Congress rule interpretations*, *Appendix B-9*. (Washington, Library of Congress, Cataloging Distribution Service). Loose-leaf, examined April 4, 2011.

3. Library of Congress, "Library of Congress will not change "dept." to "department" at this time". 1/3/11.
http://www.loc.gov/catdir/cpso/department.html

4. Joint Steering Committee for Development of RDA, *Resource description & access: RDA*. (Chicago: American Library Association, 2010).

5. Jeffrey Beall, "The Weaknesses of Full-Text Searching," *Journal of Academic Librarianship* 34, no. 5 (2008): 438-444.

6. Terada, Akira, Takenobu Tokunaga, and Hozumi Tanaka. "Automatic Expansion of Abbreviations by Using Context and Character Information," *Information Processing & Management* 40, no. 1 (2004): 31.

7. Bernard J. Jansen, Amanda Spink, and Tefko Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing & Management* 36, no. 2 (2000): 207.

8. Rob Young, "AlltheWeb search closes its doors," *Search Engine Journal*, April 5, 2011. http://www.searchenginejournal.com/alltheweb-search-closes-its-doors/29029/

9. Bernard J. Jansen and Amanda Spink, "An Analysis of Web Searching by European AlltheWeb.com Users," *Information Processing & Management* 41, no. 2 (2005): 361-381.

10. Bernard J. Jansen, "The Methodology of Search Log Analysis," in *Handbook of Research on Web log Analysis* (Hershey, PA: Idea Group Inc., 2008), 99-121.

11. Ibid, p. 101.

12. Some percentages don't add up to 100 due to rounding.