Exploring the Validity of Electronic Newspaper Databases

Travis N. Ridout
Associate Professor
Department of Political Science
Washington State University
816 Johnson Tower, Troy Lane
Pullman WA 99164-4880
United States
1-509-335-2264
1-509-335-7990 (fax)
tnridout@wsu.edu

Erika Franklin Fowler
Assistant Professor
Department of Government
Wesleyan University
238 Church Street
Middletown CT 06459-0019
United States
1-860-685-3407
efowler@wesleyan.edu

Kathleen Searles
Assistant Professor
Department of Political Science
Augusta State University
2500 Walton Way
Augusta GA 29841
United States
ksearles@aug.edu

21 August 2011

Key words: validity, electronic databases, newspapers

Word Count: 5900

Biographies

Travis N. Ridout is Associate Professor of Political Science at Washington State University (Washington, U.S.A.) and serves as co-director of the Wesleyan Media Project. His research interests include campaigns, elections and political advertising,

Erika Franklin Fowler is Assistant Professor of Government at Wesleyan University (Connecticut, U.S.A.) and serves as co-director of the Wesleyan Media Project. Her research interests include the study of local news media and political advertising.

Kathleen Searles is Assistant Professor of Political Science at Augusta State University (Georgia, U.S.A.). Her research explores the use and impact of emotional cues in political messages.

Abstract

Do electronic newspaper databases contain all of the stories that appear in the print edition? And does this depend on the database used? To explore these questions, we collected print copies of newspapers from cities across the United States and Canada. We compared coverage of two topics in these newspapers with the coverage from key word searches in three electronic newspaper databases. We conclude that the stories obtained through electronic searches are consistent across databases but can vary from the print source. However, national and international coverage is more likely to be missing than local or statewide/provincial coverage.

Exploring the Validity of Electronic Newspaper Databases

Researchers who need to analyze newspaper content almost always turn to electronic database searches nowadays. The advantages of electronic search over the "old-fashioned" method of obtaining newspapers and searching line-by-line for the subject of interest are numerous. One advantage is the ability to access newspapers from around the globe. Moreover, researchers save considerable time when they isolate articles of interest via key word searches, as opposed to scanning manually through pages of text or, for those newspapers that had them, consulting a periodical index. One could only hope that such an index was available for the newspaper you wanted—and that whoever compiled the index did a thorough job. Even then, there were problems with using such indices, such as incompleteness (Althaus, et al. 2001) and the changing meaning of words over time (Soothill and Grover 1997).

Clearly, electronic newspaper databases have made the task of the content analyst much easier. In doing so, they have made it easier for researchers to explore local media coverage instead of focusing on elite newspapers merely because they are in most libraries and have an associated periodical index. Moreover, textual analysis tool development is likely to greatly increase the extent to which scholars mine electronic databases (Monroe and Schrodt 2008). Yet research examining the validity of such databases—how well they replicate their print counterparts and each other—is scant, not to mention dated. Does it matter which electronic database one uses (i.e., will one access the same stories in Lexis-Nexis as ProQuest and NewsBank)? And are researchers who rely on electronic databases accessing the same stories that appear in the print edition? The answers have important methodological implications for researchers who utilize electronic newspaper databases for data collection.

To investigate such questions, we collected print copies of newspapers from over a dozen cities across the U.S. and Canada for two days in April 2009. We compared coverage of two

topics in the print edition with the coverage obtained from key word searches in three electronic newspaper databases. In addition, we compared the complete content from both the front-page and front ”metro” section of the print newspapers with the content available in each of the three electronic newspaper databases. Although our study is exploratory, the data collected suggest that electronically obtained stories are fairly consistent across databases but can vary considerably from the printed copy. We also find a systematic pattern to the type of missing article. In all but the largest newspapers, articles about national and international issues are more likely to be missing from the electronic databases than local or statewide/provincial stories. Thus, scholars may be underestimating the amount of national and international attention in local newspapers and consequently the amount of such coverage to which people are exposed. Our results suggest that content analysts may need to amend their data collection strategies.

Use of Electronic Newspaper Databases

Electronic newspaper databases are used for three main purposes in the social sciences. One is to obtain a record of the count of some particular event, referred to as “event count studies” (Woolley 2000), for which other systematic data are limited (the number of riots or armed conflicts in a certain area, for example) (Franzosi 1987). Another use of such databases is in studies of news media attention to an issue, referred to as “studies of media focus” (Woolley 2006). For instance, one study measured the salience of presidential candidates in the 2000 election by conducting a Lexis-Nexis search and calculating the proportion of campaign stories mentioning each candidate (Son and Weaver 2003). Another study of media agenda-setting regarding foreign affairs involved a Lexis-Nexis search to produce a monthly article count as a proxy for media attention (Soroka 2003).

In addition to the two uses outlined above, there is a third reason scholars might turn to electronic databases: to examine how journalists construct news articles. For example, Haigh and colleagues (2006) study the tone, framing and authoritativeness of embedded versus nonembedded coverage of the U.S. Iraqi invasion. Our research speaks most directly to media focus, as we are concerned less with the frequency of external events or how journalists assemble stories and more with how well we can characterize the volume and content of media coverage of a particular topic.

Newspaper Database Validity

We are interested in two types of validity associated with the use of electronic newspaper databases: (1) whether stories are identical across databases, and (2) whether electronic content matches that of the print edition. Research speaking to each of these types of validity is not extensive, and most is decades-old, dating from when electronic newspaper databases were a new tool.[1] Moreover, the most extensive study of these questions to date exists only as a conference paper (Snider and Janda 1998). Further research on this topic is warranted.

By and large, comparisons of print and electronic versions of newspapers find considerable differences (Kaufman, et al. 1994; Snider and Janda 1998). Reasons given for these inconsistencies are numerous. For one, the print edition may be a different one from that uploaded to the database. Some newspapers may provide content from all of their editions, while some may include only one. Some newspapers upload wire service and syndicated content, while others upload only staff-written and free-lance articles. Virtually no newspaper provides sports box scores, classified ads, display ads, meeting calendars, legal notices,

photographs or stock market listings. Sometimes editorials, letters to the editor, obituaries, short articles and tables will be omitted (Orenstein 1993).

Given that wire service stories are less likely to appear in the electronic database, there should be specific types of articles that are more likely to be missing, namely, those that feature national and international news. All except the largest newspapers rely upon a wire service for news outside the state/province, while almost all local news—and much statewide/provincial news, depending on the newspaper's size—is generated by local reporters. Thus, researchers who rely upon electronic searches of local newspapers may underestimate the amount of national and international coverage available to readers.

Take, for example, one study of local newspaper coverage of members of the United States' Congress (Schaffner and Sellers 2003). Based on a Lexis-Nexis search of newspapers in 40 states, the authors conclude that local newspapers give no more coverage to congressional leaders than other members. But if their searches failed to obtain much of the newspapers' wire service content, then they may have missed stories about Congressional leaders. Because it caters to readers across the United States and the world, we expect that the Associated Press is more likely to distribute articles that mention Congressional leaders than rank-and-file members—Cook (1989) found that true of national network news in the United States—but A.P. articles are more likely to be missing from an electronic newspaper database. As such, it is possible that the article's findings were an artifact of the way the sample was obtained. Our goal is not to single out one study for criticism; many researchers make the assumption that content obtained through electronic databases is a reflection of what appears in the print edition. Rather we argue for careful consideration of the methodological implications for such an assumption.

Some researchers report differences in the articles obtained depending on the database queried. For instance, one study compared the number of articles retrieved from the Dow Jones electronic database, the Nexis electronic database and Burrelle's newspaper clipping service (Snider and Janda 1998). The study discovered that the clipping service identified five A.P. articles not available in the electronic databases, but the electronic databases also identified articles that the clipping service did not. Because these were short articles, their omission was likely due to human error. In another example, Wooley (2000) found considerable differences in the amount of *Washington Post* coverage of child abuse depending on whether he used the print Washington Post Index, an electronic periodical index, the *Post*'s website archives or the Dow Jones electronic database. Another study that compared Lexis-Nexis with Google News found that the two retrieved similar articles in the New York Times, but Lexis-Nexis failed to find many articles from small newspapers that Google News retrieved (Weaver and Bimber 2008).

Overall, the limited existing research suggests that we are likely to find fewer relevant articles in a search of an electronic newspaper database than a search of the print edition. Moreover, we expect agreement between print and electronic content to vary by topic and by newspaper, though it is unclear how large those differences may be. If differences across databases or mediums are systematic, methodological approaches to studying how news is covered and its impact on audiences may need to be altered.

Data and Method

To examine the validity of electronic newspaper databases, we employed three strategies: (1) we checked the keyword search results of prominent terms against the entire content of each newspaper, and (2) we took every story from the front page and the front page of the "metro"

section of each newspaper and searched the electronic databases individually for each, and (3) we categorized the content of each one of these front page stories according to type of coverage (local, national or international) to determine whether there were any systematic patterns to missing content. Our multi-pronged strategy allows us to check both the completeness of keyword searches and to assess more systematically what types of stories appear in electronic databases. We chose to look at front page (and metro section front page) articles because placement is a prime indicator of how important a story is deemed by the publisher, and for researchers trying to assess both agenda-setting and people's exposure to the news, these stories are similarly crucial.

Our first analysis compared coverage of two different subjects (Barack Obama and the economy) across print copies of 15 North American newspapers and three electronic newspaper databases: Access World News (NewsBank), Lexis-Nexis and ProQuest Newspapers. We chose these topics because each received considerable attention, providing more observations for assessing validity. We examined coverage on Wednesday, April 15, and Thursday, April 16, 2009. Newspapers were chosen to reflect geographic diversity and variation in circulation size and were obtained from family and friends living in each city. Table 1 lists the newspapers included along with circulation sizes. All four broad geographic regions of the U. S., along with Canada, are represented in the sample. Moreover, newspaper circulation ranges from 50,000 for Michigan's *Ann Arbor News* to 1.1 million for the *New York Times*. Most newspapers are available in at least two of the databases, which yields considerable leverage when we compare the number of stories retrieved across databases.

[Table 1 here]

Our basic procedure was to search the full text of each newspaper for the dates 15 and 16 April 2009.[2] We first searched on the term "Barack Obama" with quotation marks, meaning that only those stories containing both first and last name in order were retrieved; references to "Mr. Obama" or "President Obama" would not be captured. The absence of such results from the search is not a problem because our chief interest is in the degree of fit between printed and electronic sources of news content; our human coder noted only those mentions of Obama that contained both his first and last names in order. We then recorded the number of articles obtained from the search and saved all electronic search content so that it could be referenced in the final step. We repeated this search for the term "economy," recording the number of articles obtained. The data reported here come from searches conducted on 11 May 2009.[3] Next, a coder thoroughly searched each of the print newspapers, identifying all instances of these key words. Finally, a coder searched to see if the print articles with instances of these key words were present in the electronic search content for each database.

For our second analysis, we identified all articles on the front page and front page of the "metro" section of each newspaper in our sample, finding 305 articles.[4] We then searched to see if each of these front page and metro front page articles could be retrieved in each database. Then a coder classified each of these articles as to whether or not it mentioned local issues, statewide issues in the United States/provincial issues in Canada, national issues or international issues. The coder also classified each article as to which level (local, statewide/provincial, national or international) was mentioned first. To calculate the reliability of these measures, a second coder examined each of these articles. Reliability was acceptable, with 82 percent agreement on which level was mentioned first and agreement ranging from 82 percent (the local level) to 91 percent (the international level) on individual classifications.

Finally, we supplemented these quantitative analyses with telephone interviews with a data librarian at one newspaper and representatives from two of the electronic newspaper databases.

Results

How much do the results vary across electronic databases? By and large, the differences in the number of stories retrieved were not large. Figure 1 plots in three panels the relationship between the number of stories retrieved that mention "Barack Obama" for each database pair. Figure 2 plots the same for mentions of the economy. Observations that fall on the plotted 45 degree line indicate a perfect match in the number of stories between the two databases. Most observations fell close to this line, indicating a high degree of correspondence in the number of stories retrieved across databases.

There were some exceptions to the general agreement between databases on number of stories retrieved, and this was especially true with regard to mentions of Obama. For instance, Lexis-Nexis reported 54 articles that mentioned Obama in the *New York Times* on our two search days, while ProQuest and NewsBank each reported 5. Lexis-Nexis retrieved 53 articles that mentioned Obama in the *Washington Post*, while ProQuest and NewsBank each found 6.

[Figure 1 and Figure 2 here]

The reason for these large differences is an issue of how the databases process searches. The *Times* and *Post* seldom refer to Obama by his first name, and so ProQuest and NewsBank retrieved only the small number of articles in which the token "Barack Obama" was used. By contrast, Lexis-Nexis retrieved articles that mentioned "President Obama", "Mr. Obama" or "Obama administration" even though we explicitly searched for articles mentioning his first and

last name. Indeed, if we modify the search term to "Obama," the number of stories is much more comparable across sources (61 articles for Lexis-Nexis and NewsBank and 57 for ProQuest in the *Times* and 61, 59, and 61 respectively in the *Post*). As evidenced, differences in search terms can be extremely important, but once we take into account differences in search term rules, all three databases retrieved a similar number of articles.

These findings are important for the scholar interested in the amount of attention the news media pay to a particular topic such as the economy because they suggest that regardless of the database one uses, one's measure of attention will be similar. Admittedly, however, this evidence does not speak to whether the databases identified the exact same stories. To investigate this question, we looked across all databases to see how often different stories were identified. Regardless of which two databases we compared, we found that the vast majority of articles identified were the same ones, with a smattering of articles identified by one and not the other. Agreement was 89 percent between Lexis-Nexis and ProQuest, 89 percent between NewsBank and ProQuest and 90 percent between Lexis-Nexis and NewsBank.

Print versus Electronic Sources. How fully does each electronic database cover the content of the print newspaper? A coder scanned the print copy of each newspaper, noting all mentions of "Barack Obama" and the "economy."[5] We first examine the proportion of coverage in print editions that are retrievable through electronic databases before turning to what is available electronically but not in print. Table 2 reports the proportion of print articles that mentioned "Barack Obama" in each newspaper that were identified by the electronic databases.

[Table 2 here]

There is considerable variation across newspapers in the amount of Obama coverage that made it into the electronic newspaper databases. For instance, only 13% of the *Spokesman*

Review articles that mentioned Obama made it into the databases. By contrast, all of the articles in the *New York Times* and *Washington Post* were identified. A second regularity is that the proportion of print coverage that was retrieved for a particular newspaper was generally consistent across each database. For instance, all three sources retrieved 93% of the "Barack Obama" articles in the *Toronto Star*. Finally, while there were not major differences across databases in the proportion of print coverage retrieved, ProQuest did the best, finding 56% of articles. Lexis-Nexis was close behind with 54%, and NewsBank trailed with 42%. Note, however, that some of the difference in success rate across databases may be explained by the slightly different set of newspapers within each. If we confine our analysis to only those newspapers available in all three databases, we find that NewsBank identified 43% of the print articles that mentioned Obama, and ProQuest and Lexis-Nexis each identified 48% of the Obama articles.

In general, the electronic databases did a better job of retrieving mentions of the economy. As Table 3 shows, ProQuest identified 72% of the articles, while the other two sources identified 69%. Again, however, there was variation across newspapers in the completeness of coverage. Each electronic source identified only about a third of the mentions of the economy in the Sarasota newspaper, compared to around 90% of *Boston Globe* mentions. Again, confining our analysis only to those newspapers available in all three databases paints a similar picture, with Lexis-Nexis identifying 64% of the articles the mentioned the economy and NewsBank and ProQuest identifying 68% of them. In sum, many articles in the print version edition cannot be found in electronic databases.

[Table 3 here]

We can do better explaining which articles are missing if we examine those stories' characteristics. A coder noted whether each "missing" article (article identified in print but not in the databases) was a wire service story, a news brief, a letter to the editor, a photo caption, a column, an editorial, a sidebar, or part of a special supplement. Table 4 shows the breakdown of the missing articles by type. Consider Lexis-Nexis. Of the 445 total articles on Obama or the economy identified in the print newspapers searchable through Lexis-Nexis, 283 were identified in the electronic database. Of the 162 not located (36% of the total), 86 were wire service stories. News briefs or items contained in a newspaper supplement combined constituted another 25 articles. We also identified seven sidebars that were missing, along with seven columns, six editorials, two letters to the editor, and one photo caption. The remaining 28 articles were articles written by newspaper staff or other unclassifiable content. By and large, the news articles that were contained in the print newspapers but missing in Lexis-Nexis were wire service stories, and this also holds true for NewsBank and ProQuest.

[Table 4 here]

Some newspapers in our sample uploaded all wire service stories to the electronic databases. For instance, all 12 wire service stories that we found in the print *Boston Globe* were found in Lexis-Nexis, NewsBank and ProQuest. The same is true for the 12 wire service stories we located in the print edition of the *Toronto Star*. Yet none of the wire service stories in the print editions of the Spokane, Sarasota or St. Petersburg newspapers appeared in any of the electronic newspaper databases. And then there was the *Globe and Mail*. Six of ten wire service stories appeared in the Lexis-Nexis database; eight of ten appeared in ProQuest, and none of the ten were retrieved from NewsBank.[6] Overall, only 31% of wire stories in the print editions covered by Lexis-Nexis made it into that database. The comparable percentages for NewsBank

and ProQuest are 24 and 33. In an interview, a data librarian at the *Spokesman Review* in Spokane emphasized that the newspaper does not have rights to wire content and thus does not upload it.[7] Whether other types of news content were contained in the electronic databases was unpredictable. Sometimes news briefs were included in the electronic database, and sometimes they were not. Although some photo captions were missing, it appears that some were included as separate articles in the electronic newspaper databases. Frequently, letters to the editor were retrieved, but on occasion they were not. And sometimes two or three letters that addressed the same topic might be included as a single article instead of separate articles.

The U.S. Supreme Court, in its 2001 *New York Times v. Tasini* decision, ruled that free-lance work could not be licensed by the newspaper to an electronic database. But this is open to interpretation. The Spokane data librarian reported that her newspaper did not upload letters to the editor, believing the rights to such content remained with the original author, and they use similar logic in not providing guest editorials. Our interview with NewsBank and Lexis-Nexis employees confirmed that what gets uploaded to the database varies by newspaper, and so there is no one standard. In sum, one cannot be completely certain about the types of articles that are included in each database.

That said, there is a positive relationship between the percentage of content that is included and the circulation of the newspaper. Indeed, the correlation between the two for mentions of Barack Obama is 0.76 for Lexis-Nexis, 0.76 for NewsBank and 0.74 for ProQuest. In general, the greater the newspaper circulation, the more confident one can be that fewer articles are missing. One explanation for this finding may be that larger newspapers produce more of their own content and rely less upon wire service content.

One other category of article is worth mentioning: articles found in an electronic database but not located in the print source. NewsBank located four articles that we were unable to locate in the print versions of the newspapers. All had to do with the economy. We obtained them because each contained "Barack Obama" as an indexing term, even though the exact phrase "Barack Obama" was not located within the article itself. Although scholars should certainly be aware of the possibility that searches will result in "extra" articles of this type, this appears to be a rare phenomenon that is specific to NewsBank. Thus, what is available in electronic databases is primarily a subset of articles appearing in the print editions.

***Study Two: Prominent Stories & The Content of Missing Articles*.** Having examined the validity of key word searches, we were also interested in assessing the extent to which prominent stories (those on the front page and front page of the metro section) are available through electronic databases. Whereas our first investigation focused on two prominent (albeit nonrandom) topics, this investigation takes in a wider range of topics but also allows us to say more, not about the source of the missing articles, but their content as well. To recap, this investigation involved an analysis of all 305 front page and metro front page articles that appeared in our print sample of newspapers. We categorized each of these articles according to the focus of its coverage (local, state/provincial, national and international), and then we searched for each story individually in each of the electronic databases. Given the findings of our first investigation, especially with regard to wire service stories, we speculated that articles concerning national and international issues would be more likely to be missing than articles about state/provincial or local issues, given that all but the largest newspapers rely on wire services for news beyond their geographic region.

This expectation was supported, albeit not strongly. Table 5 shows the percentage of articles that we were able to locate in each database by the first level of coverage (e.g., local, state/provincial) mentioned and by newspaper. Overall, between 83% and 90% of articles that first mentioned local coverage appeared in the electronic databases, while the percentage of articles that appeared that first mentioned state/provincial coverage ranged from 84% in Lexis-Nexis to 91% in ProQuest. There was a small drop in the percentage of articles that appeared in the databases that were classified as national coverage; the range here was 79% to 82%. The sharpest decline occurred when it came to international coverage. The percentage of articles at this level that appeared in the databases ranged from 68% in NewsBank to 71% in Lexis-Nexis.

[Table 5 here]

While not revealing dramatic differences across levels of coverage, these findings are nonetheless important, as scholars frequently note the lack of international coverage in American news media—and in local news media in particular. Graber, for instance, reports that foreign affairs coverage constitutes 21% of total coverage in elite newspapers but only 8% of coverage in non-elite newspapers (Graber 2006). These calculations, however, are based on a report that relied upon electronic database searches, which likely missed international stories in the smaller newspapers.[8] That is not to say that local newspapers provide more international coverage than elite national newspapers; in our sample, 49% of stories in the national newspapers had an international angle versus 18% of the stories in local newspapers. But, in general, relying upon electronic databases will understate the amount of international coverage that appears in local newspapers.

Discussion

Our comparison of newspaper content across electronic databases and with print editions reveals several important facts about database validity. The first is that the number of results one obtains for a particular search can vary depending on the database, but that variability is minimal, suggesting that newspapers upload the same articles to each electronic database. That said, search terms can be tricky. For one, a newspaper's stylistic conventions can have an impact on the number of articles retrieved, as we found with the *NYT*, *Boston Globe* and *Washington Post*—all three of which seldom use President Obama's first name. Moreover, even when we specifically asked for articles that contained the phrase "Barack Obama", we sometimes retrieved articles that contained only the word "Obama", indicating that different search engines employ different rules for searches.

Second, electronic databases seldom come close to matching the complete content of a printed newspaper. One reason is that many fail to upload wire service stories to the electronic databases. The degree of completeness, however, ranges considerably by newspaper. We did not, however, find that one electronic database was always better than another and therefore cannot categorically endorse one over the others. Importantly, though, certain types of articles are less likely to be uploaded to electronic databases. Articles about international issues often do not appear in the databases, which has both methodological and substantive implications. Researchers may need to alter their data collection and methodology to account for systematically missing data. Also, those interested in measuring the volume and content of information reaching average Americans in local newspapers will need to tempter their findings if electronic databases are the sole source of content. The absence of wire service stories may have even greater ramifications in the future as economic pressures force newspapers to rely

increasingly on wire service content. The upshot, then, is that the validity of studies of media focus (content) that rely upon electronic newspaper databases is seriously undermined.

When should researchers forgo the ease of utilizing electronic newspaper databases for data collection in favor of the added-validity associated with coding print copies of a newspaper? The greatest cause for concern is when the newspaper under investigation is a non-elite one that relies heavily on wire-service content, which is often excluded from electronic newspaper databases. There is also greater cause for concern when the topic under investigation is an international one (or even a national one), as news closer to home tends to be covered by local reporters and thus included in electronic databases. Finally, coding print copies of a newspaper is most worthwhile when the researcher's primary aim is knowing the exact content or focus of media coverage—such as a study of media agenda setting—but does not seem worth the effort when the researcher is more interested in how journalists craft arguments or use particular frames, which may not require that the sample of stories analyzed be one that is representative of the population.

What advice can we give to researchers who employ electronic newspaper databases? First, know what type of content is contained by the source. This is a difficult task, as there is no one guidebook that indicates what type of content each newspaper archives. We thus echo the call of Kaufman and colleagues (1993) for the production of such a source. In the meantime, close examination of a few sample searches would quickly reveal whether wire service articles or letters to the editor are included for a particular newspaper in a particular database.[9]

Related to this point, researchers need to know what they are looking for. If their main objective is to analyze the rhetoric used by staff writers in campaign stories, then an electronic search will likely be a good source of information as staff-written articles are almost always

included. If the objective is to explain variations in political learning as a function of the count of mentions of a political candidate, then an electronic database search may be problematic, as many mentions—those contained in sidebars, news briefs, and wire service stories—may be missing and this missing data may significantly skew results. It might also prove difficult to compare the count of stories about a particular candidate across newspapers, as one newspaper may upload a greater percentage of news content than another. In sum, researchers should consider how the absence of some content could affect their conclusions, both substantively and methodologically.

Finally, an initial comparison of the articles retrieved from different electronic databases may be beneficial. If the numbers match closely, then one can move forward, but if the numbers do not match, a closer examination may be warranted. Researchers might also turn to online news aggregators, such as Google News. Weaver and Bimber (2008) found Google News to be better than LexisNexis in providing wire service content printed in smaller newspapers, but they note that in 2007 Google News began to exclude some wire service content, and newspapers are able to prevent Google News from indexing their online content.[10]

The scope of our project is limited in that it does not embrace the content of newspaper websites. Comparing such content with the print copies of the newspapers—and what is contained within electronic databases—would be even more difficult than what we accomplished here. One reason for that is that some news content nowadays is web-only, never making it into the printed version of the newspaper. Second, it is difficult to identify when something was published on a website. Thus, the degree of similarity between the front page of a website and the front page of a print newspaper is something that could change hourly. Yet as web readership grows, it will be increasingly important for researchers—both those who are

interested in the effects of the news media on audiences and those interested in the content produced by reporters—to take into account this web content. Saying that, we do not want to understate the importance of our own findings, as the majority of research focuses on print newspaper content as a proxy for exposure and attention. As such, most researchers employ data collection strategies which use electronic newspaper databases and for that reason questions of validity across databases and between mediums remain of great import.

Other limitations of our research are that it is limited to only two days in April 2009 and is limited to a relatively small number of newspapers. Moreover, our sample is limited to North American newspapers, and thus our findings may not travel well to different media systems, especially those in which national, as opposed to regional, newspapers dominate. Nonetheless, given the paucity of research on the validity of electronic newspaper databases and the time-intensive nature of matching up print newspaper content with electronic content, our analysis represents an important contribution. Still, a replication of this research involving a greater number of days and newspapers—along with a true random sampling of newspapers—would surely be a worthwhile path for future research.

Although our findings point to many cautions regarding the use of electronic newspaper databases, there is also some reassuring news. All three databases contain comprehensive coverage of the *New York Times* and the *Globe and Mail*—the newspapers of record in the United States and Canada, respectively. There are few articles found in the print editions of these newspapers that are not found by electronic searches, and this holds true regardless of the database and regardless of whether the story is a local or international one.

References

Althaus, S. L., Edy, J. A., & Phalen, P. F. (2001). Using substitutes for full-text news stories in content analysis: Which text is best? *American Journal of Political Science, 45*(3), 707-723.

Cook, T.E. (1989). *Making laws and making news: Media strategies in the U.S. House of Representatives*. Washington, DC: Brookings Institution.

Franzosi, R. (1987). The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods, 20*(1), 5-16.

Graber, D.A. (2006). *Mass media and American politics*. 7th ed. Washington, DC: CQ Press.

Haigh, M. M., Pfau, M., Danesi, J., Tallmon, R., Bunko, T., Nyberg, S., et al. (2006). A comparison of embedded and nonembedded print coverage of the U.S. invasion and occupation of Iraq. *Harvard International Journal of Press/Politics, 11*(2), 139-153.

Kaufman, P. A., Dykers, C. R., & Caldwell, C. (1993). Why going online for content analysis can reduce research reliability. *Journalism Quarterly, 70*(4), 824-832.

Monroe, B.L. & Schrodt, P.A. (2008). Introduction to the Special Issue: The Statistical Analysis of Political Text," *Political Analysis* 16(4): 351-355.

Orenstein, R. M. (1993). How full is full revisited: A status report on searching full-text periodicals. *Database Magazine, 16*(5), 14-22.

Schaffner, B. F., & Sellers, P. J. (2003). The structural determinants of local congressional news coverage. *Political Communication, 20*(1), 41–57.

Snider, J. H., & Janda, K. (1998). Newspapers in bytes and bits: Limitations of electronic databases for content analysis. Paper presented at the annual meeting of the American Political Science Association. Boston.

Son, Y. J., & Weaver, D. H. (2005). Another look at what moves public opinion: Media agenda setting and polls in the 2000 U.S. election. *International Journal of Public Opinion Research, 18*(2), 174-197.

Soothill, K., & Grover, C. (1997). A note on computer searches of newspapers. *Sociology, 31*(3), 591-596.

Soroka, S. N. (2003). Media, public opinion, and foreign policy. *Harvard International Journal of Press/Politics, 8*(1), 27-48.

Weaver, D.A. & Bimber, B. (2008). Finding news stories: A comparison of searches using LexisNexis and Google News. *Journalism and Mass Communication Quarterly*, 85(3), 515-30.

Woolley, J. T. (2000). Using media-based data in studies of politics. *American Journal of Political Science*, *44*(1), 156-173.

Figure 1: Relationship between Number of Search Hits of Barack Obama and Database

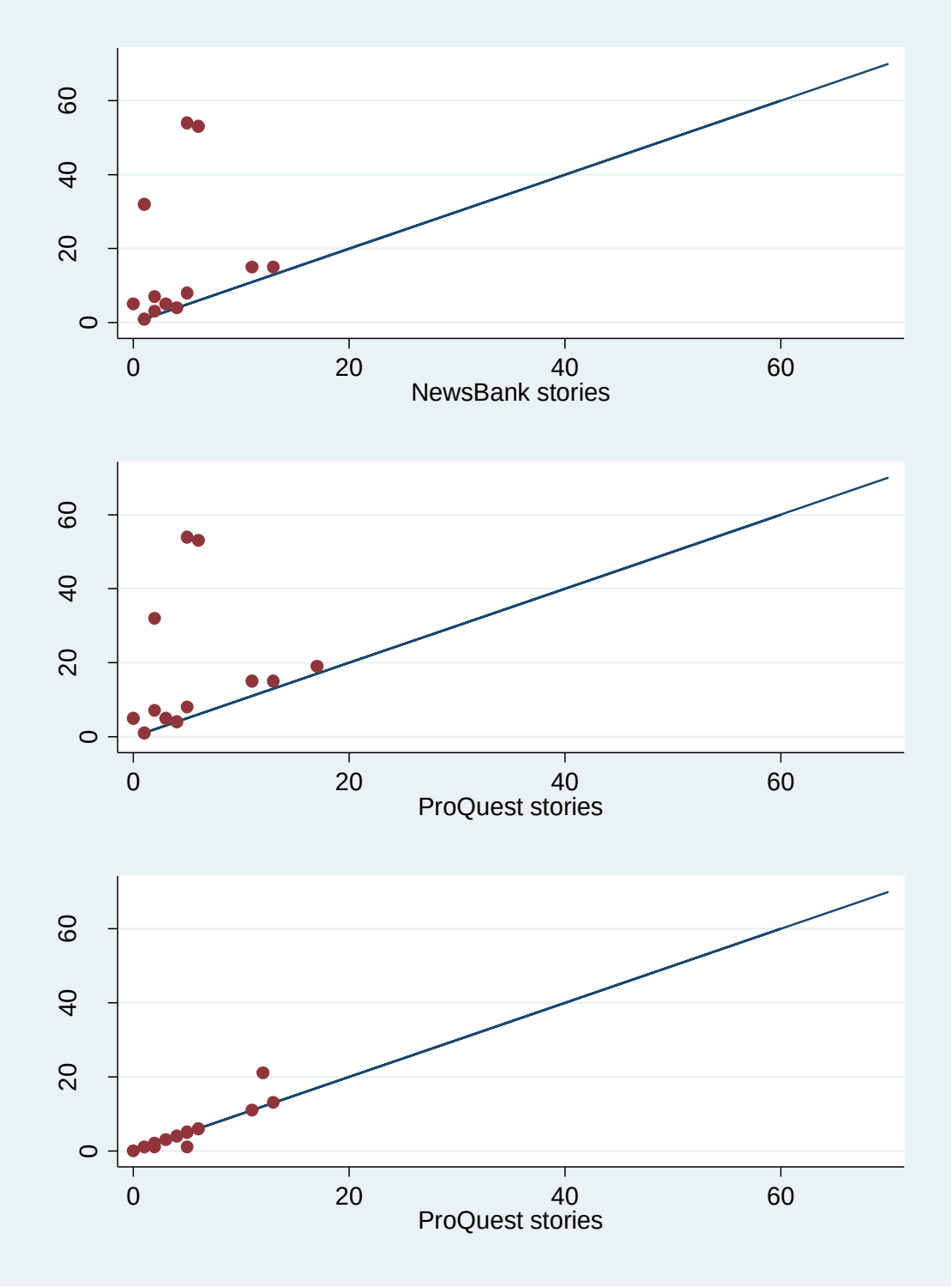

Figure 2: Relationship between Number of Search Hits of Economy and Database

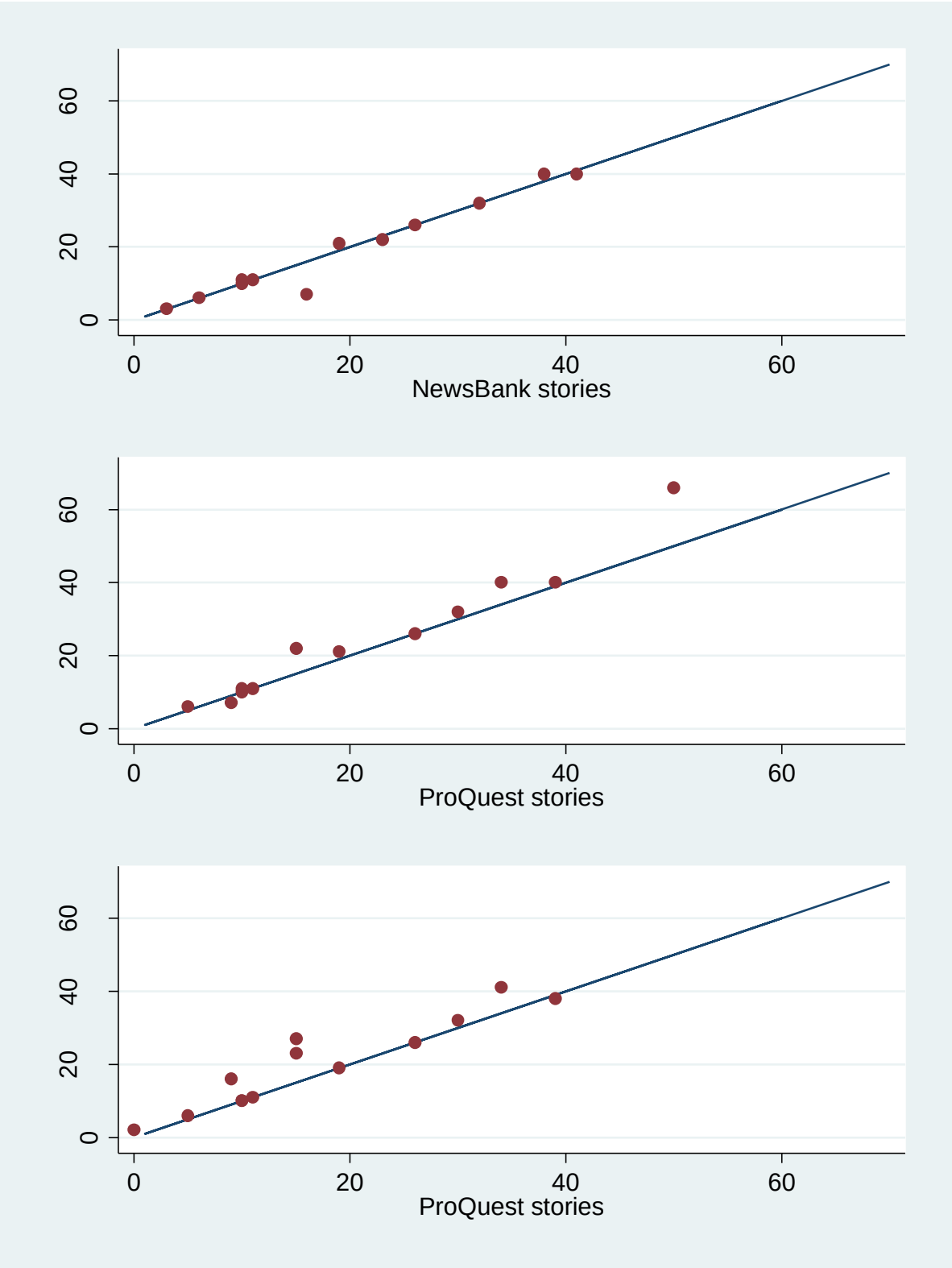

Table 1: Profile of Newspapers Contained in Sample (Sorted by Circulation)

Newspaper	Lexis Nexis	News Bank	ProQuest	Circ.	Region
New York Times	x	x	x	1,100,000	Nat'l
Washington Post	x	x	x	707,000	Nat'l
Boston Globe	x	x	x	450,000	East
Toronto Star	x	x	x	436,000	Ontario
Philadelphia Inquirer	x	x	x	368,000	East
Toronto Globe and Mail	x		x	322,000	Nat'l
Denver Post	x	x	x	275,000	West
Boston Herald	x	x	x	240,000	East
St. Petersburg Times	x	x	x	226,000	South
Detroit News		x	x	224,000	Midwest
Des Moines Register		x		152,800	Midwest
Sarasota Herald-Tribune	x	x	x	110,000	South
Spokesman Review	x	x	x	99,000	West
Wisconsin State Journal	x	x	x	90,000	Midwest
Ann Arbor News	x	x		50,000	Midwest

Table 2: Proportion of Print Articles About "Barack Obama" Found in Electronic Database Search

	Lexis-Nexis	NewsBank	ProQuest	# Articles
Ann Arbor News	0.33	0.33		6
Boston Globe	0.50	0.50	1.00	2
Boston Herald	0.67	0.67	0.67	3
Denver Post	0.29	0.06	0.29	17
Des Moines Register		0.08		13
Detroit News		0.52	0.57	23
New York Times	1.00	1.00	1.00	5
Philadelphia Inquirer				0
Sarasota Herald-Tribune	0.31	0.31	0.23	13
Spokesman Review	0.13	0.13	0.13	8
St. Petersburg Times	0.77	0.69	0.69	13
Toronto Globe and Mail	0.79		0.89	19
Toronto Star	0.93	0.93	0.93	14
Washington Post	1.00	1.00	1.00	6
Wisconsin State Journal	0.19	0.19	0.19	21
Total	0.54	0.42	0.56	163

Table 3: Proportion of Print Articles About "Economy" Found in Electronic Database Search

	Lexis-Nexis	NewsBank	ProQuest	# Articles
Ann Arbor News	0.57	0.57	0.00	7
Boston Globe	0.94	0.97	0.88	32
Boston Herald	0.91	0.73	0.91	11
Denver Post	0.52	0.52	0.48	21
Des Moines Register		0.14		14
Detroit News		0.91	0.68	22
New York Times	0.98	0.98	0.73	41
Philadelphia Inquirer	0.28	0.00	0.36	25
Sarasota Herald-Tribune	0.33	0.33	0.30	27
Spokesman Review	0.67	0.67	0.67	18
St. Petersburg Times	0.80	0.80	0.72	25
Toronto Globe and Mail	0.89		0.93	55
Toronto Star	0.90	0.90	0.90	29
Washington Post	0.39	0.95	1.00	38
Wisconsin State Journal	0.27	0.33	0.33	15
Total	0.69	0.69	0.72	380

Table 4: Number and Characteristics of Articles Not Located

	Lexis-Nexis	NewsBank	ProQuest
Number of articles from print edition	445	443	474
Number located in electronic source	283	258	317
Proportion located	.64	.58	.67
Characteristics of articles not located			
Brief	12	20	16
Caption	1	1	1
Column	7	10	10
Editorial	6	6	6
Letter	2	4	3
Sidebar	7	9	8
Supplement	13	12	13
Wire Service	86	105	85
Other	28	18	15

Table 5: Proportion of Print Articles Located by Newspaper, Source and Level of Coverage

Newspaper	Local Coverage			State/Provincial Coverage			National Coverage			International Coverage			Paper Average
	LN	NB	PQ	LN	NB	PQ	LN	NB	PQ	LN	NB	PQ	
Ann Arbor News	0.92	0.92		1.00	1.00		1.00	1.00		0.50	0.50		0.89
Boston Herald	1.00	1.00	0.80				0.00	0.00	0.00	0.00	0.00	0.00	0.67
Detroit News		0.93	0.50		0.67	0.67		0.57	0.57				0.56
Sarasota Herald-Tribune	0.43	0.86	0.79	0.50	0.67	0.67	0.67	0.67	0.67	0.00	0.00	0.00	0.64
St. Petersburg Times	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00				0.98
The Boston Globe		1.00	0.86		1.00	1.00		1.00	1.00		1.00	0.00	0.95
The Denver Post	0.83	0.83	0.83	1.00	0.88	1.00	0.00	0.00	0.00	0.50	0.50	0.50	0.76
The Globe and Mail	1.00			1.00			1.00			1.00			1.00
The New York Times	1.00	1.00	1.00	0.67	0.67	0.67	1.00	1.00	1.00	1.00	1.00	1.00	0.95
The Philadelphia Inquirer	0.53	0.65	0.65	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.67
The Spokesman-Review	0.83	0.89	0.89	1.00	1.00	1.00	0.67	0.67	0.67				0.86
The Washington Post	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.83	0.83	1.00	1.00	1.00	0.92
Toronto Star	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wisconsin State Journal	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50	0.50				0.96
Average	**0.83**	**0.90**	**0.88**	**0.84**	**0.89**	**0.91**	**0.82**	**0.79**	**0.81**	**0.71**	**0.68**	**0.70**	

Notes

[1] Proquest began distributing electronic material on CD-ROM in the 1980s and moved to electronic distribution in 1996. Lexis-Nexis began providing desktop access to select newspapers using personal computers in 1979; web-based distribution began in 1994 and catered to legal professionals. It is unclear when NewsBank began distributing newspaper content electronically.

[2] These dates were chosen for convenience but seemed fairly unremarkable in terms of news content. The big national story in most newspapers was Tea Party protests across the country. The big international story was piracy off the coast of Somalia.

[3] This does raise the issue of whether one's results are consistent over time. For another project, we repeated these searches on various days and found very little variation at all depending on the search day. Indeed, we got the exact same number of hits in when we repeated the search on 18 May as on 11 May.

[4] The *Boston Herald* does not have a "metro" section per se, as it is a tabloid. We therefore searched for articles from the first and second pages of the newspapers. At the time we conducted this analysis, the *Des Moines Register* was no longer available in any of the three databases.

[5] To be consistent with our electronic searchers, the coder did not note mentions of "Mr. Obama", "President Obama" or other variations on his name in the print copies of the newspapers.

[6] Each of the electronic databases allows one to search wire service stories separately, which may be helpful in determining whether a newspaper uploads wire service content to the database.

[7] Interview conducted via email on 21 October 2009.

[8] Graber relies upon data from *The State of the News Media 2004* report, produced by the Project for Excellence in Journalism.

[9] Quite simply, since approaches to uploading content vary from newspaper to newspaper and from database to database, we mean that the researcher would need to sit down and scan all of the results retrieved to see whether any letters to the editor, editorials, photo captions, wire service stories (i.e, those types of content often excluded from an electronic database) appear or not.

[10] Google. News (publishers) help. "Restricted Content." http://www.google.com/support/news_pub/bin/answer.py?hl=en&answer=68331. Accessed on August 16, 2011.