Newspapers in Bytes and Bits:

Limitations of Electronic Databases for Content Analysis

J.H. Snider and Kenneth Janda

Northwestern University
Evanston, Illinois 60208

Prepared for delivery at the 1998 Annual Meeting of the
American Political Science Association,
The Marriott Hotel, Boston,
September 3-6, 1998

How closely do electronic databases of newspapers reflect print versions? This question grows in importance as scholars exploit the convenience of electronic databases for content analysis of social and political documents. Our study produces questions and some answers that raise cause for concern.

Several years ago a prestigious communications journal rejected an article that depended for its results on a library search through the contents of a physical publication—a publication readily available on NEXIS. The reviewer explained that the results of the article were less valid and reliable than they would have been had the researcher used an electronic database. The reviewer further suggested that the researcher redo the analysis from the ground up using an electronic database. Might the reviewer have been wrong? Might the burden of proof have more appropriately rested on the research design using the electronic database? More generally, how much confidence should we have in the reliability and validity of electronic database searches?

Defining the Databases for Study

The term, "electronic database," has been used in various ways. Most people agree on the minimal definition of a "database" as a collection of persistent, related information.[1] This information, according to the World Intellectual Property Organization (WIPO), includes "collections of literary, musical or audiovisual works or any other kind of works, or collections of other materials such as texts, sounds, images, numbers, facts, or data representing any other matter or substance [including] .... expressions of folklore."[2] So any durable collection of related material constitutes a database.

An *electronic* database is defined to include capacity for automated search and retrieval. One standard source describes it as "a collection of files used to store information that is managed by a database management system, or DBMS."[3] The everyday meaning of "electronic database," however, depends on the user's profession. To most librarians, electronic databases are bibliographic files.[4] But to most quantitatively oriented social scientists, an electronic database holds large files of numbers that can be deciphered with corresponding documents, called codebooks. To a smaller but growing group of political scientists and scholars in communications studies, an electronic database can mean vast bodies of natural language text converted to digital codes and stored in searchable files. This genre of electronic databases concerns us here.

By "natural language text," we mean the writings and recorded speech of both elites and ordinary citizens. In outline form, here are specific examples, organized by source of text in this partial listing:

I. The public as a source of natural language text:
   - A. Comments to open-ended questions on interview schedules
   - B. Recorded discussions in focus groups
   - C. Letters to public officials
   - D. Letters to the editor

II. Governmental sources in the U.S.
   A. Presidency (e.g., inaugural addresses, executive orders, news conferences)
   B. Congress (e.g., floor debate, testimony in committee hearings)
   C. Supreme Court (e.g., opinions on cases)

III. Extra-governmental political institutions as sources
   A. Platforms of political parties
   B. Statements and advertisements issued by candidates for election
   C. Transcripts of debates between candidates
   D. Statements and advertisements issued by interest groups

IV. Media sources
   A. Television & radio (e.g., transcripts of news programs)
   B. Newspapers & magazines (e.g., editorials, opinion columns, news reports)

Electronic databases have been created for virtually all of these categories of natural language text. We will focus on the last source, media reports of social and political events contained in newspapers. As described in Appendix A, our findings about electronic databases of newspapers probably also apply to electronic databases of television and radio news transcripts. Media sources probably account for the largest body of natural text in electronic form, and they are being used increasingly in content analysis on social and political topics.

Content Analysis in Social and Political Research

Most social scientists have heard about content analysis, but few have actually used the methodology extensively in research. Despite the pioneering work in content analysis by Lasswell and Pool,[5] later political scientists appear to have used it less frequently than sociologists and far less than scholars in journalism and communications studies. Due to the many ways to conduct content analysis, it is difficult to corral in a widely-accepted definition. Shapiro and Markoff examine six well-known definitions that vary in the way they treat the symbolic objects, intellectual products, nature of inference, degree of quantification, and other aspects of the method. Their own "minimal" definition of content analysis is "any methodological measurement applied to text (or other symbolic material) for social science purposes."[6] In the next sentence, they elaborate their definition to include "any systematic reduction of a flow of text (or other symbols) to a standard set of statistically manipulable symbols representing the presence, the intensity, or the frequency of some characteristics relevant to social science."

Long before computers, content analysis was done manually and produced important findings.[7] But it was a laborious process and very hard to replicate. Scholars turned to computers to assist in the analysis of natural language text in the late 1950s, even before computers had the capability to produce lower case letters and common symbols of punctuation, such as the semicolon and question mark.[8] Although computers greatly sped the analysis once text was rendered machine-readable, computers then offered no help with the human process of recording the text to analyze at machine speeds. Content

analysis remained a technique plagued with high-costs relative to its rewards. Now that most text originates in digital form and existing documents can be scanned and converted into computer files, the cost/benefit ratio of content analysis is more favorable for social research.

Content analysis has become a central tool of communications research. A 1997 review found that 34.8% of the articles published during 1995 *by Journalism and Mass Communication Quarterly* (JMCQ) used content analysis (Riffe and Freitag 1997). From 1971 to 1995 JMCQ published 486 full-length articles using content analysis, 24.6% of the total 1,977 research articles published. Of the media analyzed in the JMCQ survey, 46.7% were for newspapers and 24.3% for television. Other scholarly publications addressing political communications, *including Political Communication, Journal of Communications*, and *the Harvard Journal of Press/Politics*, also make frequent use of newspaper and television content analysis.

Although we have gathered no data to support the claim that the use of electronic databases for such content analysis is increasing, it is a fair presumption that it is. One indicator is the growth in the amount of mass media available for electronic content analysis. According to BiblioData, the number of fulltext online sources increased from about 3,000 in 1989 to about 43,000 in 1998 (Fulltext Sources Online 1998). Dow Jones News Retrieval (hereafter "Dow Jones") increased its coverage of *fulltext* daily United States newspapers from 1 in 1980 to 113 in 1997 (see Figure 1). Its total coverage of all daily U.S. newspapers (including abstracts and select text) increased from 1 in 1969 to 288 in 1988. In the early 1980s academics at top United States universities would be lucky to have twenty daily United States newspapers accessible in their university library. Thanks to the information revolution, by the late 1990s they can expect to have more than ten times as many.

Enter Figure 1

At the same time, the cost of using electronic databases has plummeted. More university libraries offer scholars free and unlimited access to such databases as LEXIS-NEXIS (hereafter "NEXIS"), Dow Jones, and ProQuest Direct. Online vendor competition is greater than ever before. In 1989 each fulltext source was carried by two different vendors on average. By 1993, the figure rose to 4.6 (" 1993). Vendors' adoption of the Internet and standard browser interfaces make searching easier than ever before. Many databases can now be accessed over the Internet without even entering a library.

Assuming that the online newspaper databases accurately mirror the print publications, the arguments for using the news articles in electronic form would seem compelling. Electronic databases can save researchers huge amounts of precious time and money. One author noted that a content analysis that took over 500 hours in 1986 could now be done in a few seconds (Keith and Grover 1997). The need for teams of researchers and paid assistants can be radically reduced.

Classic problems of validity and reliability are attenuated or even eliminated. Instead of searching through a small sample of newspapers and articles, scholars can now peruse hundreds of newspapers and thousands of articles, thus reducing problems of statistical

inference from samples to populations. Instead of analyzing language with simple rules that human brains can easily remember and quickly apply, fast and consistent electronic processors can use complex algorithms, thus allowing for more sophisticated and accurate content analysis. Instead of relying on error-prone human beings to count the articles and analyze them, error-free computers can be relied upon, thus eliminating problems of reliability.

Certain important problems once deemed impractical to investigate now seem doable. For example, the availability of hundreds of online newspapers could help bridge the gap in the political communication literature between national and local media. Traditionally, scholars have overwhelmingly focused their analysis on national media such as the *New York Times, Wall Street Journal*, and *Washington Post*. But the vast majority of Americans do not read such papers. Not only were studies of local newspapers hard to do, but their validity could be questioned based on the necessity of using a small sample of the country's more than 1,500 daily newspapers. Consequently, much less is known about local than national media, although it can be argued that in many important political spheres local media are much more influential and important.

Literature Review

Unfortunately, using electronic databases for content analysis can have severe drawbacks because of the common discrepancy between print and electronic versions of publications. These drawbacks were noted almost as soon as fulltext databases became common. As early as 1987, Pagell, a librarian at the University of Pennsylvania, asked "how full is full?" After comparing the results of hardcopy and online searches, Pagell concluded that the answer depends on the system and database, "but the glass is never filled to the brim" (p. 36). By 1993, the online search community had written numerous articles about the undocumented omissions in online databases. Orenstein, revisiting Pagell's question six years after her original article, concluded "that all that Pagell said six years ago is still true, only more so (1993, p. 14).[9]

In 1993, Kaufman et al. made the first serious scholarly attempt to look at the discrepancy between print and electronic versions of newspapers. Their "most basic finding was that content analysis performed on a database yielded results different from the hand search" (p. 826).

In 1995, Hansen concurred with Kaufman et al.'s results and provided additional insight into why such discrepancies happen. She concluded her brief paper with the observation that "content analysis that relies on the electronic data base version of a newspaper is seriously compromised" (p. 5).

Nevertheless, Kaufman et al. and Hansen are rarely cited in research that employs content analysis.. The large survey of content analysis cited at the beginning of this paper (Riffe and Freitag 1997) is indicative of the inattention paid to problems of validity and reliability associated with using electronic databases. The article applauded the growing use of sophisticated methods to enhance the validity and reliability of content analysis, but none of the methods involved care in the use of electronic data Other articles, some

specifically focused on the methodological problems of doing newspaper content analysis, have similarly ignored problems of electronic data (e.g., Lacy et al. 1995; Riffe et al. 1993).

Since the pathbreaking work of Pagell (1987) and Kaufman et al. (1993), availability and use of online newspapers has exploded. NEXIS and Dow Jones have become staples at many university libraries, and scholars are doing more online newspaper searches than ever before. At the same time, electronic transcripts of TV and radio news programs have become readily available. Problems in the use of such transcripts have been ignored in the scholarly community and have received barely more attention in the community of professional online searchers.

Here we provide an up-to-date look at concerns political communication scholars should be aware of when they do electronic content analysis of mass media. Specifically, we provides an overview of the limitations of electronic databases for analyzing the content of newspapers and, to a lesser extent, television. We provide information not only on the exclusions but also insight into why those exclusions exist, which ones can be relatively easily controlled for, and which ones are likely to be relatively permanent as opposed to short-term glitches based on flawed technology and systems.

Unlike many articles on electronic databases, the focus here is on issues of validity and reliability rather than convenience or cost. Of course, the database researcher should not ignore such issues in choosing among both vendors and individual vendor offerings. Vendors differ in the way they allow searching and printing of records. They also typically charge a premium for convenient searching and printing options. For example, one vendor may only allow one newspaper article at a time to be viewed and printed. Another may allow twenty. Such variations can significantly affect the amount of time a search will take. Since many vendors charge for time online, increasing searching and printing speed can cost them money. Vendors, wary of copyright theft, also have incentives to make downloading of data inefficient. Scholars wading through thousands of articles should take such concerns very seriously.

One reason for downplaying such issues is that vendor interfaces and pricing are continuously changing. For example, in the last few years the Internet has led the online vendors to offer completely revamped interfaces. Anything we write here would likely be soon outdated. In contrast, as we will argue, the issues of validity and reliability focused on here are likely to be more enduring.

Much of the research here relies on the published and online literature of the various electronic databases. In addition, Snider interviewed numerous customer service representatives at Burrelle's, NEXIS, Dow Jones, and Dialog; marketing department contacts at Burrelle's, NEXIS, Dow Jones, Dialog, and ProQuest; Susan Bjorner, author of *Newspapers Online*; Ruth Ornstein, author of *Fulltext Sources Online*; Nora Paul of the Poynter Institute; Marydee Ojala, editor of *Database*; Barbara Quint, editor of *Searcher*; and electronic rights specialists at the Associated Press and New York Times.

Availability of Fulltext Newspapers

Many political communication scholars consider NEXIS to be the gold standard for doing online content analysis of newspapers. Yet as we shall see, not only is NEXIS no longer dominant in newspaper coverage, but the whole notion of a gold standard, including one-stop shopping for all online newspapers, needs to be re-examined.

In Table 1 we report in detail how thoroughly the major newspaper data bases cover the top 100 daily newspapers in the United States, according to circulation. These newspapers represent approximately 71% of total daily United States newspaper circulation (Editor & Publisher 1997). In their product catalogs, database vendors usually describe their databases as falling into one of three categories: **F**ulltext, **S**elected fulltext, and **A**bstracts. Following vendor terminology, Table 1 tags each database's coverage as F, S, or A.

Enter Table 1

In Figure 2, we summarize and graph the data in Table 1 according to the two categories of newspaper archives: print and electronic. Within each category, the sources are ranked roughly in order of claimed thoroughness.

Enter Figure 2

Print databases: Burrelle's, the country's leading clipping service, comes out on top for thoroughness among the databases that hold the physical newspapers. As of April 1997, Burrelle's claimed to cover 1,676 daily United States papers, and it had all of our top 100. Burrelle's is also the most convenient way to do a paper-based search through a large number of newspapers. A brief telephone call to a Burrelle's representative is all that is needed. Burrelle's is usually hired on a long-term basis to track mentions of companies, products, individuals, and issues. Burrelle's has a lag time of about a week between the time a newspaper is published and read for clipping, so it does allow for at least some retrospective searching. Researchers can wait to get a sense of how a story is playing before hiring Burrelle's.[10]

Among libraries, the Library of Congress (LOC) comes out on top, with 350 daily United States newspapers on microfilm, including the top 100. If you are searching a large number of newspapers, it is advisable to telephone the LOC's newspapers and periodicals room ahead of time. Describe your search and request permission to view all the desired microfilms at once. Otherwise, you will only be able to request up to two microfilms at a time after 12 p.m. and four at a time before 12 p.m.

Of course, the newspaper holdings of most libraries will not compare with the Library of Congress. For reference, consider Northwestern University's newspaper microfilm collection. It has extensive holdings of daily papers, but only collects microfilm on 15 of the top 100 daily newspapers.

Electronic databases: Of the electronic database vendors, Dow Jones comes out on top, followed closely by NEXIS. Dow Jones has 96 of the top 100 daily newspapers online; NEXIS has 95. Dow Jones has 78 of the top 100 newspapers in fulltext; NEXIS has 76.

The data indicate it is possible to do fulltext searches on the top 59 daily newspapers in the United States, but only by using Dow Jones, NEXIS, and Dialog in tandem. The combination of Dow Jones and Dialog will allow searching on 58 of the top 59—or all 59 if only the last 90 days of the New York Times are needed. Of the top 100 daily newspapers, only 7 are not available in fulltext and only one is not available in any form whatsoever.

Underlying many of the differences in fulltext coverage are long-term strategic considerations. For example, Dow Jones publishes the *Wall Street Journal* (rank#1) and retains exclusive control over its distribution, only allowing the other database vendors to provide abstracts. In contrast, the New York Times (rank#3) was for many years exclusively available in fulltext on NEXIS. Now it is also available to other vendors, but other vendors only have rights to archive it in fulltext for 90 days whereas NEXIS archives go back more than two decades.[11] Similarly, Knight-Ridder, a large newspaper chain, owned Dialog from 1988 till 1997. A consequence is that Knight-Ridder owned newspapers such as the *Philadelphia Inquirer*(rank#16), *Detroit Free Press*(rank#20), *San Jose Mercury News*(rank#31), *Charlotte Observer*(rank#43), *St. Paul Pioneer Press*(rank#56), and *Philadelphia Daily News*(rank#61) are available in fulltext on Dialog but not on either Dow Jones or NEXIS.[12]

Another important area where vendors differ is in the depth of their archives. One vendor that carries fulltext of a particular newspaper may begin coverage later than another. For example, fulltext coverage of the *San Francisco Chronicle* starts January 1985 for Dow Jones, January 1988 for Dialog, and October 1988 for NEXIS (Fulltext Sources Online 1998).

Fulltext versus Cover-to-Cover

Vendors frequently use the terms fulltext, selected fulltext, and abstract, but rarely define them. In the *1998 LEXIS-NEXIS Directory of Online Services*, a 510 page book with the density of a telephone directory, there is a 181 page section that describes every NEXIS database as fulltext, selected fulltext, or abstract. Nowhere is there a description of exactly what those terms mean. Similarly, the *1996 Publications Directory* for Dow Jones (the print directory was discontinued and replaced by an online directory in 1997) uses the same terms with no explanation. According to a different Dow Jones publicationFull text

> means that all of the editorial content of the publication is included online. Selected full text means that not all articles are included online. Abstract means that summaries of the articles are presented; these may not necessarily include every article in the publication.[13]

The terms abstract and selected fulltext are straightforward and cause little confusion, but the term "fulltext" is misused. According to Monica Sluyter, Product Manager for Dow Jones Interactive Publishing, this definition appears to be a mistake. She states that the official (and legal) definition of fulltext for Dow Jones is the more restrictive one used by BiblioData, publisher of *Fulltext Sources Online*.[14] The BiblioData definition of fulltext also appears to be the standard one used by all the major online vendors:

> One would like to think that a journal or newspaper described as being available online in fulltext is available cover-to-cover. This is rarely the case.... In no case do online periodicals reproduce advertisements that appear in the original. None, to our knowledge, reproduces long tabular material such as pages of stock quotations. Because of copyright restrictions, most newspapers exclude wire service stories and syndicated columns online, limiting themselves to items written by in-house staff. In addition, many omit letters to the editor, editorials, obituaries and filler material (Fulltext Sources Online 1998, p. v).

BiblioData classifies all such periodicals as "fulltext." The term "selected" fulltext is reserved for periodicals with even more restricted inclusion criteria. For example, Business Dateline, a periodical collection including several hundred daily newspapers (found in Dialog, Dow Jones, and NEXIS), "limit the number of articles taken from any particular publication to those that are business news." (Fulltext Sources Online 1998, p. v).

The term fulltext may seem more reasonable in a historical context. It originally evolved to contrast with the pervasive use of article abstracts by online databases. The term fulltext meant the fulltext of an article, not the fulltext of a publication.The importance of the incorrect Dow Jones definition of fulltext is that it does capture a vendor tendency to publicize fulltext inclusions rather than exclusions. Vendors use the term fulltext in a context that implies complete coverage of a publication from cover-to-cover. Even those who search out vendor qualifications of the term "fulltext" may find themselves confused. Vendor written descriptions of exclusions are incomplete and ambiguous at best; misleading at worst. According to Bob Simons, retired counsel for Dialog Corporation, ambiguous exclusion statements are a way for newspapers and vendors to keep their options open in a world where contracts change frequently and public documents can be a source of legal liability.[15] Consider the NEXIS description of the *Boston Globe*'s exclusions. The entire description is one sentence.

> EXCLUSIONS: All articles from syndicated or freelance writers.[16]

Does this include newswires? Are AP articles included? Does it include articles from the New York Times, the parent company of the *Boston Globe*? How about the *Boston Globe*'s own syndicated columnists? What about other types of common exclusions such as multiple editions, letters-to-the-editor, retractions, and magazine sections (often not written by staff)? How have the *Boston Globe*'s policies regarding exclusions changed since NEXIS coverage of the *Boston Globe* started in January 1987?[17]

It appears that the *Boston Globe*'s official policy is to exclude all newswire services and syndicated columns, even from its own parent company, the New York Times. For example, a William Safire column, distributed by the New York Times Syndicate, should not appear in the NEXIS file of the *Boston Globe*.[18]

Some exclusions are of little importance to political communication scholars. In this category we would include financial data (e.g., stock prices, interest rates, mutual fund prices, and money exchange rates), vital records (e.g., birth announcements, engagements, and wedding announcements), obituaries, court records, crime reports, community

listings, entertainment listings, games & puzzles, sports statistics, weather statistics, graphics & photos, display advertising, classified advertising, and comics.

Other exclusions, relating to coverage of news and political opinion, can be very important. These include wire service stories, letters-to-the-editor, op-eds, magazine sections, retractions, zoned editions, timed editions, and changed policies regarding all of the foregoing. It may be the peculiar misfortune of political communication scholars that the exclusions most likely to affect the validity and reliability of their findings are the least well documented.

Finally, fulltext does not even mean the complete content of an article. Graphics, captions, tables, and inserts can all be missing from an article that is still labeled "fulltext." In addition, visual information, such as size and placement of headlines, is not currently captured in online databases (Soothill and Grover 1997).

Important Exclusions

Retractions. Some newspapers retract articles that become sources of controversy or litigation. Ojala, editor of *Database* magazine, describes the widespread use of retractions. Most of her examples come from medical publications, but she also describes one newspaper, the *Virginia-Pilot*, that retracted statements made in four different news stories plus a column about a loan made by a local sheriff candidate to a bank (Ojala 1996). In a telephone interview, Ojala described a similar retraction by the *Cincinatti Enquirer* concerning a series of articles on Chiquita Banana in 1998.[19]

Corrections. Corrections may be handled inconsistently. Some corrections are made to the original printed text. Chamberlain found that 11 of the 21 news librarians that responded to her survey said that such corrections were sometimes or usually made (1998). To the follow-up question "If so, do you note that you've done so?" she found that 7 of the 8 respondents said yes (1998). In other words, no notice may be given that the online version of an article has been purposefully changed to no longer reflect the printed version.

Multiple Editions. Information on multiple editions is often omitted. Most major newspapers have more than one zoned and timed editions, but only the edition of record is usually uploaded to database vendors. Bjorner reports that the *New York Times* includes articles in regional editions under its final Late Edition of the *Times*. It would be easy for a researcher to confuse a story that only was in the Connecticut edition of the *Times* with one in the Manhattan edition of the *Times* (Bjorner 1996, p. 36). Ingebretsen and Lutgen describe the more than 12 editions of the *Los Angeles Times* and the varying dates at which electronic coverage became available for each edition (1991, p. 18). This type of detailed data is not available from the database vendors.

Policy Changes. Coverage of newswires may vary over time, but not be reported by database vendors. Nora Paul, the former news librarian at the *Miami Herald*, said the *Herald* stopped uploading AP articles for at least several years during the early 1980s. The reasoning was that the AP articles were already available from AP, so it was not

worth the effort for the newspaper to duplicate the AP's archival storage. Marydee Ojala said the *Kansas City Star* stopped uploading non-staff articles in the early 1990s, then changed policies several years later. To her knowledge, all of this remains purely oral history.[20]

Syndicated Columnists. Most national syndicated columnists retain electronic rights. According to Vince Price at UMI, this includes authors such as George F. Will, William Safire, Otis Pike, Dave Barry, Molly Ivins, Mona Charen, Thomas Sowell, and Clarence Page.[21] However, columnists have the ability to wave electronic rights for any particular article that they write.

Newswires. The most important exclusion for political communication scholars is probably newswires. According to one study of a Gannett chain newspaper, the number of wireline written stories exceed staff-written pieces (Coulson and Hansen 1995). Since the vast majority of daily newspapers in the United States depend on the wire services as the foundation for their national news, the absence of the wire services would make electronic databases virtually useless for tracking local coverage of national stories. For example, a researcher tracking local coverage of telecommunications policy or the Monica Lewinsky scandal would come out largely empty handed. The exception would be the major newspaper groups and their flagship papers, such as the *New York Times*, *Washington Post*, *Chicago Tribune*, *Los Angeles Times*, and *USA Today*, which generate most of their own national news (Ojala 1997). According to David Tomlin, General Executive of the Membership Department at the Associated Press, "Newspapers have no rights to redistribute AP material. AP contracts only specify that newspapers have rights to publish AP stories in the print versions of their papers. They do not mention electronic rights." The AP does allow newspapers to purchase AP material for limited time periods for use in their new, Internet-based editions (they even allow them to use AP material not printed in the newspaper), but this does not apply to AP material uploaded to database vendors for resale. Such use is strictly forbidden.[22]

In gathering information about exclusions for this paper, database vendors provided the most inconsistent information about exclusions related to electronic rights and newswires. A half dozen calls to NEXIS service representatives seemed to generate six different theories about how newswires were treated. Even senior officials at the top database vendors and newswires appeared to be confused.

The actual situation appears to be as follows. Like the Clinton Administration's policy of no gays in the military, there is a de facto "don't ask; don't tell" policy regarding newspaper usage of newswires and syndicated columns. Wire services for the most part retain electronic rights to articles but are not diligent about enforcing whether their customers adhere to them. Local newspapers, which can rely on wire services as a foundation for most of their national and even state news, are not anxious to cut this information from their electronic archives. The only people who seem to know the details of the policies are the corporate counsels who negotiate the confidential contracts. Companies might not want to publicly acknowledge widespread violations of their contracts--even to their own employees--because it might put them in a difficult legal situation sometime in the future.

In any case, the newswires may not lose much revenue from the current policies. NEXIS automatically deletes AP articles within newspapers files from library and group file searches. The official explanation is that this reduces duplicate articles. Most professional searches do not want to get 50 versions of the same AP article. If they get articles directly from the AP feed, this is adequate. An additional advantage of this relationship is that the AP gets the revenue from the search. When the AP article comes from the newspaper file, the revenue goes to the newspaper.[23] However, researchers can get around this problem by doing searches directly on particular newspaper files, something most political communication researchers will want to do anyway but that most commercial researchers probably do not do.

Genuine Errors

Another type of reason for the discrepancy between print and electronic editions has to do with errors. News librarians, those responsible for uploading newspaper articles from the newspaper to the online vendor, seem to be most concerned with this type of discrepancy. In the only major study of this type of problem, Oakley found these errors to be common (1997; 1998). Others have also examined the problem (Semonche 1993; Chamberlain 1998).

Errors can and do frequently occur at any point in the information flow from reporter to page designer to news librarian to database vendor. Presently, the weakest link in the information flow appears to be the gap between the editorial and design side of newspapers. After a writer submits an article for publication, the design side can change headlines, cut text, or even cut an article entirely. Depending on whether the computers on the design and editorial side are integrated, these changes may or may not be reflected in the final version of the article uploaded to a database vendor. Oakley tells the illustrative story of Jack Brummett, a political columnist at the *Arkansas Democrat-Gazette*, who wrote a false and libelous statement about a public figure. Alerted to the problem, he went to the page designer and got the error fixed in the page proofs, but the original version was uploaded to the database vendor. The database error brought legal action by the public figure's attorney. In this case, the error was corrected. But most errors appear not to be either discovered or corrected. He quotes a news librarian who says "that there is no way for her small staff to take about 100 articles a day and 'edit them line by line and get it right.'" (1998, p. 13). Chamberlin quotes a news librarian who says a "newsroom was supposed to make sure the electronic content was the same as the print content--but never did. Are we surprised?" (1998).

Some problems are caused by computer glitches, not human error. Chamberlin found a software problem that led articles from some newspapers to be truncated after they were transmitted to NEXIS. Specifically, long stories that had corrections appended or that were resent with fixes became truncated on retransmission. Any files sent before the software patch was developed may still be corrupt (Oakley 1997). Oakley's advice to newspapers also applies to political communication scholars:

> Do you assume the archival capture comes after final page proof corrections? Better check. Assume headlines, captions, and corrections

> are electronically cut and pasted where they belong rather than retyped before archiving? Better check. Assume corrected versions sent to a commercial database supplant incorrect originals? Better check. Assume that an article retrieved from a commercial database matches newsprint? Better double check (Oakley 1998, p. 13).

Reliability Problems

Researchers should not necessarily expect that similar searches will generate similar results. One scholar reported that after he published an article he went back to NEXIS to check the article counts. The results were different (chances are he was searching on a group file, the component files of which often change).

The basic reliability problem is that online databases are constantly changing and these changes are not documented. Even when the changes are documented , it can be a great challenge to keep up with them. Simply specifying an exact search string (or more than one search string if different vendors are used) is not enough to ensure that other scholars can replicate results. All the underlying databases would have to be described in depth, including their structure (e.g., fields), content (e.g., exclusions), differences across vendors (e.g., in uploading policies), and changes (e.g., in structure, content, and uploading policies).

As for relying on vendors' article counts, scholars must be careful to devise their own definition of article and then investigate whether the articles found qualify. Snider has had as many as five identical articles show up in the same publication, perhaps because of an idiosyncratic way of dealing with multiple editions. Some long articles with separate boxes get sliced and diced into three or four distinct articles. Scholars should not necessarily expect consistent definitions of articles across publication, vendor, or time. Newspapers rendered in bytes are often in pieces.

Comparison of Print vs. Electronic Newspaper Coverage

To investigate the relationship between print and electronic coverage of news, we designed the following test. Burrelle's newspaper clipping service was hired to clip all AP articles by the Associated Press on the subject of digital television from April 2 through April 8. In particular, Snider was interested in tracking the distribution of a series of articles by AP telecommunications writer Jeannine Aversa. The articles concerned the FCC's April 3, 1997 decision to grant additional spectrum to broadcasters for digital television. The assumption was that Burrelle's readers, who use printed versions of newspapers to do their searches, would provide a reference point to evaluate the thoroughness of the fulltext searches on the online databases. The two online databases chosen for comparison were Dow Jones and NEXIS. These were chosen for their comprehensive coverage of top newspapers as well as their availability in academic libraries.

The top 52 daily newspapers were chosen because Dow Jones has a convenient group file of the top fifty United States newspapers by circulation (in actuality, the file has 52

newspapers). As discussed above, NEXIS files were searched individually to avoid the problem of group files excluding AP articles

Table 2 shows the results. The Burrelle's search generated 21 AP articles, the Dow Jones search 36, and the NEXIS search 30. The Burrelle's results were quite surprising. Burrelle's is widely considered to be the leading clipping service in North America. A New York Times executive had recommended Burrelle's as the organization the New York Times would use if it wanted to check on who was carrying its specific syndicated columns.

Enter Table 2

Judith Mandelbaum of Burrelle's offered two reasons Burrelle's may have missed articles: human error and different editions. The edition Burrelle's uses may not be the edition an online service uses. This results in discrepancies. Burrelle's also relies on human readers, and they sometimes miss articles.[24]

A review of the missing articles suggests Burrelle's readers are especially likely to miss short articles (e.g., a paragraph in length). Some newspapers do not properly credit the AP for AP-derived articles. These articles were not credited against Burrelle's. However, one very brief story (only three sentences in length)credited the AP within the body of the story (i.e., "the Associated Press reported") rather than in the byline. Another story gave the byline to "Bee News Services" and had an acknowledgment at the end of the article that read "Sources: USA Today research, Associated Press." Both articles were credited against Burrelle's.

Burrelle's did locate five articles that neither Dow Jones nor NEXIS located. This is consistent with the hypothesis that not all AP articles are uploaded by newspapers to online services.

The high number of articles for both Dow Jones and NEXIS suggests that AP-derived articles are widely available online. It would appear that the great majority of daily United States newspapers have no fear about being sued for uploading AP material.

Also interesting is the discrepancy between NEXIS and Dow Jones counts. Dow Jones found six more AP-derived articles than NEXIS. Explanations for four of the six discrepancies can be inferred reasonably easily: Omission #1: The *Chicago Tribune* ran two articles on the FCC digital TV decision. One was an AP article; the other was by its Washington correspondent, Frank James. The James article ran in one edition; the AP article in another. Dow Jones picked up the AP edition; NEXIS the James edition. Omissions #2 and #3: NEXIS only carries selected fulltext of the *Portland Oregonian* and *San Antonio-Express News*. In contrast, Dow Jones carries the fulltext of those two publications. Omission #4: NEXIS missed a tiny article in the *Rocky Mountain News*. The same paper had another, larger AP story that both NEXIS and Dow Jones picked up.

One point of claimed superiority by Dow Jones did not seem to be warranted. Dow Jones claims to have selected "Fulltext" of the *Philadelphia Inquirer* whereas Nexis only claims to have an abstract. Nevertheless, Dow Jones and Nexis both pulled up the same one

sentence abstract (32 words in length) of a *Philadelphia Inquirer* article. Moreover, the abstract concerned a front page article with important business implications, exactly the type of article one would expect to find in a selected fulltext database. If such an article can be missing from a database asserting to be selected fulltext, then even seemingly modest claims to selected fulltext should be viewed with skepticism.

The results suggest that Dow Jones has a slight edge in comprehensive coverage of the top 52 daily newspapers. They suggest that online newspapers can be a good source of local newspaper articles about national news--even if the local papers do not clearly have electronic rights to the articles they upload. They also suggest that reliance on NEXIS and Dow Jones alone will likely miss articles found in the hardcopy or microfilm versions of the newspaper.

Conclusion

Database vendors have minimal incentive to publicize exclusions. Most customers cannot afford to subscribe to or use more than one database vendor. Customers place a high premium in having all the publications they want available on one vendor. Not surprisingly, all the vendors claim to be the largest data source of fulltext information.

As long as both librarians and scholars rely on vendor information, problems of validity and reliability will be hard to avoid if online databases are used. The problem is aggravated by the incentives of both wire services and newspapers to obfuscate their policies and practices regarding electronic rights to news material.

By restricting searches to narrow classes of information--editorials and local, staff-produced news--scholars can confidently use online services with minimal knowledge of the exclusion policies of the particular newspapers under investigation. Unfortunately, this strategy precludes the study of national issues that have traditionally been the focus of investigation by political communication scholars. By restricting searches to newspapers that produce most of their own national news (newspapers such as the *New York Times*, *Wall Street Journal*, *Washington Post*, and Los Angeles), scholars can also be fairly confident of minimal omissions. An important caveat is that these newspapers also are less likely to upload the few outside stories they do buy.

Scholars should be more willing to use more than one source of newspaper information. Even if hard copies of newspapers are ultimately used, online searches can be extremely helpful in doing a pilot test and identifying promising issues and articles. Given the enormous effort and expense required to do a hard-copy search, it would seem foolish in this day and age not to take advantage of electronic resources.

Those scholars attempting to do a large-scale study should be wary of relying too heavily on a single online vendor. Vendors have complementary strengths and are willing to pay a premium for exclusive control of certain highly sought databases. In the future, it is possible that online vendors, which are essentially middlemen, will be bypassed by newspapers that provide powerful search capabilities directly to their customers. Already, newspapers are making major efforts in this direction.

Similarly, it is possible that wire services will one day completely bypass newspapers. Reuters, in particular, has made major strides in this direction. Reuters is already a major source of direct national and international news via the Internet, usually accessed via search engines such as Yahoo and Excite.

Scholars who do employ multiple vendors for searching must be wary about inconsistencies. Different databases employ different definitions of articles (e.g., some will treat boxed material as separate articles), search syntax (e.g., for NEXIS the field name to describe an author is "byline;" for Dow Jones it is "source"), and field content (e.g., some vendors include subtitles in headlines; others only include the main title). However, it must be recognized that searches on a single online vendor may suffer from many of the same problems. Online vendors often bring together very different databases under one umbrella. The simple search interfaces they present to the public can be quite misleading if the underlying databases differ substantially.

Problems with manual searches should not be overlooked. The cost of checking for and compensating for problems of validity and reliabilty may be as prohibitively expensive for manual searches as they are for electronic searches. For graduates students and others without large grants, online searches may be the only practical alternative to do important, if flawed, research.

This paper has overlooked the problem of electronic editions of newspapers (see Martin and Hansen 1996; Martin and Hansen Forthcoming). Until a few years ago, an online newspaper almost always meant an archive of a paper newspaper. Today, it is necessary to distinguish between electronic *archives* of newspapers and electronic *editions* of newspapers. Electronic editions of newspapers are likely to gradually supplant paper-based newspapers. Such newspapers already widely differ from their paper-based cousins and are likely to increasingly differ in the future. The issue of the discrepancy between the paper and electronic version of a newspaper becomes obsolete in such a context. Indeed, the very notion of an archive becomes problemmatical in a web-based world where links become evanescent and "papers" continuously change.

This paper has only touched on the massive growth of electronic transcripts of TV and radio news (see appendix A). The growth of electronic transcripts has been more recent and dramatic than the growth of online newspapers. But many of the problems associated with online newspapers also apply to online transcripts. Unlike online newspapers, however, the biggest hurdles to the development of useful TV and radio transcripts may not relate to the sins of the marketplace but to the sins of government policy, especially regarding copyrights. Although the government has granted TV and radio broadcasters billions of dollars in government subsidies (including free spectrum, subsidized towers, and favorable tax laws) it has asked for little in return in terms of accountability and copyright law. Unlike most copyright holders, broadcasters are not required by law to submit copies of their works to the Library of Congress. They have won onerous fair use restrictions on their news programs and generally do not allow the public to see their archives. It is arguably a great misfortune for American democracy that the most powerful political media in the United States--local TV news programs--have no public record and therefore minimal accountability. In contrast, it is fairly easy to test newspapers for bias and other

sins and then hold them accountable. It would be a fairly easy request to ask broadcasters to submit their teletext versions of newscasts and/or the feeds from their teleprompters to the Library of Congress. With the advent of high quality speaker-dependent software, the process of transcript creation could even be automated if the broadcasters willingly went along.

Perhaps the majority of national news uploaded by daily newspapers is done illegally--i.e., without explicit electronic rights to do so. This may appear shocking, but a little reflection reveals that such arrangements are pervasive in relations between government and private industry and in relations among private companies. Many laws are not enforced, and so are many contracts between companies. National news sources retain their electronic rights to keep their options open for the future, but they see little to gain and more to lose by enforcing those rights right now. To be fair, there is still uncertainty about the contractual claims of the electronic rights holders. According to Bob Simons, retired corporate counsel for Dialog Corporation, it is possible that their claims will not be upheld in court.[25]

What scholars need to keep in mind is that the accuracy of online newspaper archives is very much in flux. The situation today may be very different from the situation in the year 2000. For all the problems with electronic archives outlined in this paper, it is possible we are nevertheless living in a scholarly golden age when electronic rights are not enforced and valuable information is easily accessible.

In conclusion, it cannot be overemphasized that the services provided by vendors change on an almost daily basis. Certain types of problems, such as the widespread errors in uploading newspaper information, are likely to be solved in coming years. Other types of problems, such as the discrepancy between print and electronic rights, are more enduring and will likely become even worse. This paper has focused on some of these more enduring problems. They should be the focus of much ongoing research in future years.

Appendix A:

Electronic Databases of Television News Transcripts

In recent years, electronic transcripts from more than 200 TV and radio programs have become widely available from commercial vendors. Many of the problems pertaining to online newspaper coverage also pertain to online transcript coverage.

In Table A we compare coverage of major television news programs based on different types of information sources. The information sources are divided into two categories: verbatim transcripts and abstracts. Among the vendors of verbatim transcripts, none has comprehensive coverage. Burrelle's is noteworthy for having many transcripts with the earliest start dates. NEXIS stands out for its exclusive coverage of CNN programs and its slightly earlier coverage of many ABC programs. Dow Jones is noteworthy for the great expansion of its coverage in recent years.

Among all venders, Vanderbilt Archives is unique in the depth of its coverage. It covers the weekday evening news for ABC, CBS, and NBC all the way back to August 5, 1968. Video Monitoring Services (VMS), available through NEXIS and partly owned by Burrelle's principals, is unique for the breadth of its coverage. Most notably, it covers local TV news for the top 42 United States markets. This includes coverage of every evening news program in each market. Most VMS coverage only begins after January 1993. VMS uses the abstracts to sell video clips, for which it charges $105 per five minute clip. Video Monitoring Services tapes 60,000 hours per month. They retain local news tapes for 31 days and national news tapes for 60 days.

Little appears to be known about the validity and reliability of TV and radio transcripts. For example, it is not generally known whether transcripts are verbatim accounts or cleaned up accounts of news shows. According to Marydee Ojala, CNN transcripts are cleaned up. If the anchor misstates a fact or gets a name mixed up, the transcript shows what was intended as opposed to what was said. Current transcripts on network websites often append the phrase "This is an unedited, uncorrected transcript." Since spoken language is often ungrammatical and poorly phrased, the temptation to clean up the spoken word must be great. For example, congressional floor speeches have traditionally been cleaned up, sometimes with substantial revisions, before publication in the *Congressional Record*. The same could happen with broadcast transcripts.

Another concern is that when people talk over each other their comments can be hard to transcribe. News programs such as the MacNeill/Lehrer Newshour, where people do not talk over each other, may tend to be transcribed more accurately (Ojala 1991, p. 39). In any case, scholars must be very careful about doing proximity searches on transcripts. The problem is that people interrupt each other on TV news programs, so that a sentence may be broken in the middle and not continued till hundreds of words later.

Perhaps the greatest concerns have to do with thoroughness. A vendor that claims coverage from a certain date does not necessarily guarantee 100% coverage of every segment aired by that program. Snider did a search in VMS for programs on digital TV. Of the 215 radio and TV news programs retrieved, two had portions marked "bad audio."

Both of the bad audio segments involved *TV* programs. One wonders how frequently "bad audio" segments go unmarked.

TV and radio stations may also suffer from the same problem as newspapers in not having rights to retransmit on different media their non-staff produced material. Many TV stations make heavy use of stringers and other independent sources of video. According to Lutzker, an intellectual property lawyer for the broadcasting industry, "unless a written agreement with a stringer defines his or her contribution as a 'work made for hire' or specifically embraces reuses and resale to other media, those rights may not be held by the media" (Lutzker 1997, p. 101). For example, one company in Los Angeles uses a helicopter to trail traffic accidents on major thoroughfares, and it licenses this footage to the local TV stations. After the Rodney King Verdict, the helicopter got exclusive footage of the Reginald Denny beating. It gave rights for local uses to the local TV stations, but retained rights to royalties from all other media. These rights were then sold separately to the national news media (Lutzker 1997, p. 103).

In regard to services that provide abstracts, one must always be concerned about the quality and consistency of the people doing the abstracts. Although abstracts have faded in importance for newspaper research, they are still essential for pre-1990s (the Vanderbilt Archives) or local (VMS) TV news research.

---

[1] "For instance, a filing cabinet containing one or more folders of tax returns for the last 10 years is a form of a non-electronic database that is used to store historical tax information." http://personalweb.edge.net/~spock/database.htm

[2] James Love, "A primer on the proposed WIPO treaty on database extraction rights that will be considered in December 1996," Consumer Project on Technology, Center for Study of Responsive Law (CSRL), http://www.essential.org/cpt, Revised November 10, 1996.

[3] C.J. Date, *An Introduction to Database Systems, Fifth Ed.* (1991).

[4] For example, the home page of the Humbolt State University Library in the California system says:

> Each record in a data base is composed of the important elements of information for a particular item. For example, in Periodical Abstracts the information about a single periodical article is a record. Each record is composed of a set of fields which contain the individual elements of information. For example, each record in the Periodical Abstracts database includes the fields: title, author, source, and descriptors(http://library.humboldt.edu/library/infoservices/literacy/module3/dbasdef.htm).

[5] Growing out of research conducted for the U.S. Government in World War II, Harold D. Lasswell, Nathan Leites, and associates involved in that research published the landmark book, *Language of Politics: Studies in Quantitative Semantics* (New York: George W. Stewart, 1949). A related work is Ithiel de Sola Pool, et al. *The Prestige Press: A Comparative Study of Political Symbols* (Cambridge: M.I.T. Press, 1970).

[6] Gilbert Shapiro and John Markoff, "A Matter of Definition," in Carl W. Roberts (ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (Mahwah, NJ: Lawrence Erlbaum Associates, 1997), pp. 11-14.

[7] See Frank L. Klingberg, "The Historical Alternation of Moods in American Foreign Policy," *World Politics* (January, 1952), 239-273, for a mind-boggling, manual content analysis of material on our foreign policy for more than a century.

[8]Stone, Philip J. et al. *The General Inquirer: A Computer Approach to Content Analysis: Studies in Psychology, Sociology, Anthropology, and Political Science*. Cambridge, Mass.: The M.I.T. Press, 1966.

[9] Much of the best literature on the limitations of electronic databases can still be found in occasional articles in such publications as *Database*, *Online*, *Searcher*, *Editor & Publisher*, and *Cyberskeptic*. *Newspapers Online* (Bjorner 1995) provides the most thorough available description of the exclusions for daily newspapers online but it is unfortunately discontinued due to lack of demand.

[10] Readers interested in hiring a clipping service are advised to consult with Luce Press Clippings as well as Burrelle's.

[11] No such restrictions exist on access to the *New York Times* outside the United States and Canada.

[12] In 1997 Data Times was purchased by UMI and is now included as part of UMI's ProQuest service. Subsequently, UMI developed a close working relationship with Dow Jones, vaulting both into the same league as Nexis in newspaper coverage. ProQuest newspaper coverage is similar to Dow Jones, but lacks access to several Dow Jones papers including the *Los Angeles Times* and the *Globe and Mail*.

[13] "Comparison of Dow Jones Interactive and Nexis in Coverage of Top Vertical Market Publications." New York: Dow Jones & Company Interactive Publishing News, January 28, 1998.

[14] Telephone conversation with Monica Sluyter on June 12, 1998.

[15] Telephone interview with Bob Simons on July 31, 1998.

[16] The *Boston Globe*'s exclusions may actually be a model of clarity. Consider the *Arizona Republic*'s exclusions (downloaded July 24, 1998): "Licensor may exclude some articles at their own descretion; i.e. syndicated columnists." Note the "i.e." A more accurate characterization would undoubtedly be "e.g."

[17] *Newspapers Online, 3rd Edition* (Bjorner 1995) has a more detailed description of *Boston Globe* inclusions and exclusions. The 3rd Edition explicitly states that the *Boston Globe* uploads AP articles. Assuming that the 3rd Edition is accurate, It would appear that the *Boston Globe*'s policies have changed since late 1994 when the last edition of *Newspapers Online* was compiled.

[18] To clarify such concerns, Snider asked NEXIS, Dow Jones, and Dialog to send him either the *Boston Globe*'s contract provisions regarding exclusions or their standard boilerplate contract language for exclusions in daily newspapers. All refused, saying that contractual information was proprietary. Snider also called both the *Boston Globe*'s news librarian and permissions director. Both did not return his phone calls. A former news librarian told Snider that standard newspaper contracts include a non-disclosure clause. In other words, only vendors and newspapers in litigation with each other can disclose detailed contractual exclusions.

[19] Snider telephone interview with Marydee Ojala on July 27, 1998.

[20] Snider telephone interviews with Nora Paul on July 27, 1998 and Marydee Ojala also on July 27, 1998.

[21] Telephone interview with Vince Price on July 2, 1998. Snider's experience searching for William Safire columns suggests that his columns are widely available online, regardless of whether he retains electronic rights.

[22] Snider telephone interview with David Tomlin on July 28, 1998.

[23] Scholars suffer an additional problem. Many academic subscriptions exclude the AP wire because it is too costly. The result is that when searching on library or group files, they neither get the AP article from the AP wire nor the AP article within the newspaper. This little exclusion cost Snider many hours to figure out. It was not in the published NEXIS product catalog which both he and the university librarian used as a reference. The librarian assumed Snider was doing something wrong, and Snider, depending on the published material, thought he must be retarded (a feeling he often has when using online services).

[24] Snider telephone interview with Judith Mandelbaum on July 17, 1998.

[25] Snider telephone interview with Bob Simons on July 31, 1998.

References

Bjorner, Susanne. 1996. "Where in the World is the New York Times?" *Database*, June/July, pp. 28-40

Chamberlain, Jackie. 1998. "Survey of Selected Fulltext Online Newspaper Database Quality Control Policies and Procedures." April. www.sunsite.unc.edu/journalism/oaktre.html. Unpublished.

Coulson, David C. and Anne Hansen. 1995. "The Louisville Courier-Journal's News Content After Purchase By Gannett." *Journalism & Mass Communication Quarterly*, 72(spring):205-215.

Fulltext Sources Online. 1998. Needham, Massachusetts: BiblioData, January.

Hansen, Kathleen A. 1995. "Online Inaccuracies: The Use and Misuse of Electronic Information Sources." Presented at the AEJMC Annual Conference, August.

Ingebretsen, Dorothy. 1991. "The *Los Angeles Times*: A Special Kind of Database." Database Searcher, May, pp. 17-30. [to show how history is not included; LA Times editions have come online at different times; all the vendors I examined made no such notices]

International Directory of News Libraries, Fifth Edition. 1996.Bayside; New York: LDA Publishers.

Kaufman, Philip A., Carol Reese Dykers, and Carole Caldwell. 1993. "Why Going Online for Content Analysis Can Reduce Research Reliability." *Journalism Quarterly* Winter. 70(4):824-832.

Lacy, Stephen, Kay Robinson, and Daniel Riffe. 1995. "Sample Size in Content Analysis of Weekly Newspapers." *Journalism & Mass Communication Quarterly*, 72(summer):336-345.

Martin, Shannon, and Kathleen A. Hansen. *Newspapers of Record in a Digital Age: From Hot Type to Hot Links*. New York: Praeger. Forthcoming.

Martin, Shannon, and Kathleen A. Hansen. 1996. Examining the"Virtual Publication As A "Newspaper of Record." *Communication Law and Policy* Autumn 1(4): 579-594.

Oakley, Bruce W. 1997. "Accuracy in Electronic Archives: An Investigation," April. www.sunsite.unc.edu/journalism/oaktre.html. Unpublished.

Oakley, Bruce W. 1998. "How Accurate Are Your Archives?" *Columbia Jornalism Review*, March/April, p. 13.

Oakley, Bruce William. 1998. "How Accurate Are Your Archives?" *Columbia Journalism Review*, March/April.

Ojala, Marydee. 1991. "Online Broadcast News: From Television Screen To Computer Screen." *Database*, April. pp. 33-40

Ojala, Marydee. 1997. "The Personality Characteristics of Newswires." *Database*, April/May.

Orenstein, Ruth. 1989. "The Fullness of Full Text." *Database Searcher*, September.

Orenstein, Ruth. 1993. "How Full is Full' Revisited: A Status Report on Searching Fulltext Periodicals. *Database*, October.

Poynder, Richard. 1998. "Lexis-NEXIS: Past and Future." Online & CD-ROM Review. 22(2):73-80.

Riffe, Daniel, and Alan Freitag. 1997. "A Content Analysis of Content Analysis: Twenty-Five Years of Journalism Quarterly Autumn." *Journalism & Mass Communication Quarterly* 74(3):515-524.

Riffe, Daniel, Charles F. Aust, and Stephen Lacy. 1993. "Constructed Week Sampling in Newspaper Content Analysis," *Journalism Quarterly*, 70(spring):133-139.

Soothill, Keith, and Chris Grover. 1997. "A Note on Computer Searches of Newspapers." *Sociology*, 31(August):591-596.

Wray, Ricardo, Kimberly Maxwell, and Robert Hornik. 1998. "Validation of On-Line Searches of Media Coverage: An Approach to Evaluation With an Example of Reporting Domestic Violence." Presented at the International Communication Association Annual Conference, July.

Figure 1:
Dow Jones News Retrieval

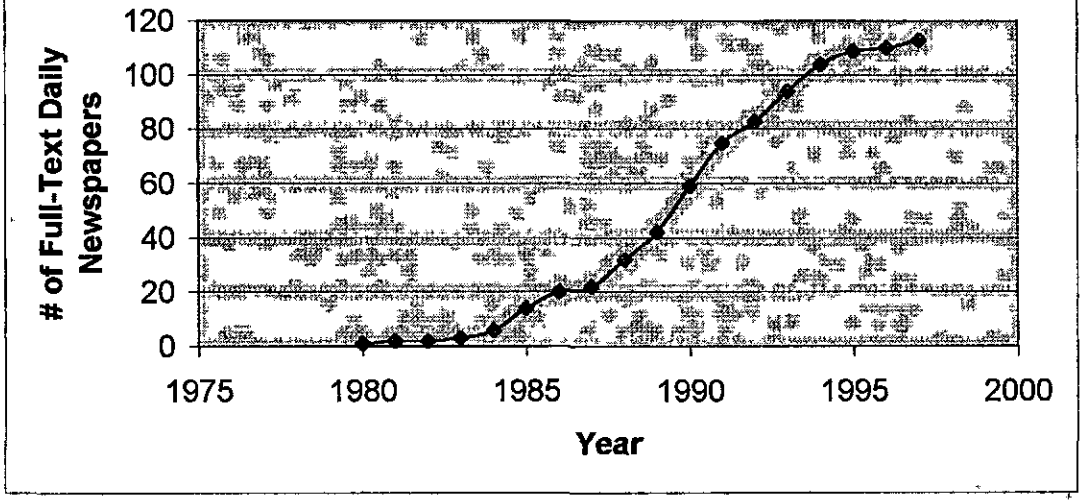

Table 1
Top 100 Daily U.S. Newspapers in Databases
(by Circulation)

			Online Newspaper Sources				Paper-Based Newspaper Sources		
Rank	Daily U.S. Newspapers	Circulation[26] as of 9/30/97	Dow Jones	Nexis	Dialog	Reuters	Burrelle's	Library of Congress	Northwestern University
1	Wall Street Journal	1,774,880	F	A	A	A	F	F	F
2	USA Today	1,629,665	F	F	F	A	F	F	F
3	New York Times	1,074,741	F	F	F	A	F	F	F
4	Los Angeles Times	1,050,176	F	F	F	F	F	F	F
5	Washington Post	775,894	F	F	F	A	F	F	F
6	New York Daily News	721,256	F	F	S	F	F	F	
7	Chicago Tribune	653,554	F	F	F	F	F	F	F
8	Newsday	568,914	F	F	F		F	F	
9	Houston Chronicle	549,101	F	F	F	F	F	F	F
10	Chicago Sun-Times	484,379	F	F	A	A	F	F	F
11	San Francisco Chronicle	484,218	F	F	F	A	F	F	
12	Dallas Morning News	481,032	F	F	S	F	F	F	
13	Boston Globe	476,966	F	F	F	F	F	F	F
14	Arizona Republic/Phoenix Gazette	437,118	F	F	F	F	F	F	
15	New York Post	436,226	A	F			F	F	
16	Philadelphia Inquirer	428,233	S	A	F	F	F	F	F
17	Newark Star-Ledger	406,010	F	A	F	S	F	F	
18	Atlanta Journal & Constitution	405,545	F	F	F	F	F	F	F
19	Minneapolis-St. Paul Star Tribune	387,412	F	F	F	A	F	F	
20	Detroit Free Press	384,624	S	A	F	F	F	F	
21	Cleveland Plain-Dealer	383,586	F	F	F	A	F	F	F
22	San Diego Union-Tribune	375,598	F	F	S	S	F	F	
23	Orange County Register	356,520	F	F	S	F	F	F	
24	Miami Herald	351,432	S	A	F	F	F	F	F
25	Portland Oregonian	342,454	F	S	F	S	F	F	
26	Denver Post	337,372	F	F	F	F	F	F	
27	St. Petersburg Times	321,447	F	F	F	F	F	F	
28	St. Louis Post-Dispatch	313,594	F	F	F	F	F	F	F
29	Baltimore Sun	312,826	F	F	F		F	F	
30	Denver Rocky Mountain News	302,953	F	F	F		F	F	
31	San Jose Mercury News	290,811	S	A	F	F	F	F	
32	Milwaukee Journal-Sentinel	288,173	F	F	F	S	F	F	
33	Sacramento Bee	281,471	F	F	F	F	F	F	
34	Boston Herald	277,106	F	F	F		F	F	
35	Kansas City Star	276,349	F	F	F	S	F	F	
36	Buffalo News	262,095	F	F	F	A	F	F	
37	New Orleans Times-Picayune	260,552	F	F	F	A	F	F	
38	Orlando Sentinel	250,886	F	F	F	F	F	F	
39	Detroit News	246,638	F	F	S	A	F	F	
40	Columbus Dispatch	246,095	F	F	F	A	F	F	
41	Pittsburgh Post-Gazette, Sun-Telegraph	243,024	F	F	F	F	F	F	
42	Fort Lauderdale Sun-Sentinel	240,091	F	F	F	S	F	F	
43	Charlotte Observer	239,016	S	A	F	F	F	F	
44	Investor's Business Daily	234,596	F	F		F	F	F	
45	Fort Worth Star Telegram	229,701	F	F	F	F	F	F	
46	Louisville Courier-Journal	228,185	F	F			F	F	F
47	Tampa Tribune	227,570	F	F	F	A	F	F	
48	Seattle Times	227,162	F	F	F	F	F	F	
49	Omaha World-Herald	225,761	F	F	F	A	F	F	
50	Indianapolis Star	224,372	F	F	S	F	F	F	
51	San Antonio Express-News	216,232	F	A	S	F	F	F	
52	Hartford Courant	210,800	F	F	A	A	F	F	
53	Richmond Times-Dispatch	209,690	F	F	F	F	F	F	
54	Oklahoma City Oklahoman	204,376	F	A,F	S	F	F	F	
55	Los Angeles Daily News	201,669	F	F	F		F	F	
56	St. Paul Pioneer Press	200,275	S	A	F	F	F	F	
57	Seattle Post-Intelligencer	197,921	F	S	F	S	F	F	
58	Cincinnati Enquirer	194,328	F	F	S	A	F	F	

			Online Newspaper Sources				Paper-Based Newspaper Sources		
59	Austin-American Statesman	178,643	F	F	F	S	F	F	
60	Rochester Democrat & Chronicle, Times-Union	177,950	A	A	A	A	F	F	
61	Philadelphia Daily News	175,290	S		F	F	F	F	
62	Memphis Commercial Appeal	174,938	F	F	F	A	F	F	
63	Florida Times-Union	171,644	F	F	S	F	F	F	
64	Arkansas Democrat-Gazette	170,766	A	F	A	A	F	F	
65	Providence Journal-Bulletin	170,292	F	F	S	F	F	F	
66	Des Moines Register	164,912	F	F	S	A	F	F	
67	Riverside Press-Enterprise	162,551	F	F	S	A	F	F	
68	Tulsa World	162,186	F	F	S	F	F	F	
69	Palm Beach Post	159,923		F	F		F	F	
70	Dayton Daily News	159,072	F	F	F		F	F	
71	Las Vegas Review-Journal	158,441	F	F			F	F	
72	Asbury Park Press	156,821	S	F	S	F	F	F	
73	Raleigh News & Observer	153,408	F	F			F	F	
74	Fresno Bee	152,718	F	F	F	F	F	F	
75	Birmingham News	150,346	A	A	A	A	F	F	
76	Syracuse Post-Standard/Herald-Journal	149,584	F	A	S	S	F	F	
77	Nashville Tennessean	146,914	F	F			F	F	
78	Record	146,089	F	F	F		F	F	
79	Toledo Blade	145,800	A	A	A	A	F	F	
80	Akron Beacon Journal	145,055	S	A	F	F	F	F	
81	Grand Rapids Press	138,907	F	A	A	A	F	F	
82	Chicago Daily Herald	132,090	F	F			F	F	
83	Salt Lake City Tribune	129,836	F	F	F	F	F	F	
84	Allentown Morning Call	128,581	F	F	F	S	F	F	
85	Tacoma News Tribune	128,498	F	F	F	F	F	F	
86	Wilmington News Journal	123,869					F	F	
87	Columbia State	121,699	S		F	F	F	F	
88	San Francisco Examiner	120,856	F	F	F	A	F	F	
89	Spokane Spokesman-Review	116,391	F	F	S		F	F	
90	Knoxville News-Sentinel	115,264	F	F			F	F	
91	Albuquerque Journal	113,694	F	F	A	A	F	F	
92	Lexington Herald-Leader	112,139	S		F	F	F	F	
93	Worcester Telegram & Gazette	108,769	F	F			F	F	
94	Charleston Post & Courier	108,637		F			F	F	
95	Madison State Journal, Capital Times	107,597	F	F	F	S	F	F	
96	Jackson Clarion-Ledger	104,375	A	A	A	A	F	F	
97	Long Beach Press-Telegram	104,078	S		S	F	F	F	
98	Honolulu Advertiser	102,236		A	S		F	F	
99	Roanoke Times & World News	102,173	F	F	F		F	F	
100	Washington Times	101,169	F	F	F	S	F	F	

Key: F=Full Text; S=Selected Full Text[27]; A=Abstracts.

Source: Compiled data from Dow Jones (current as of 2/3/98), NEXIS *(current as of 7/14/98), Northwestern University (current as of 4/8/96), Burrelle's (current as of 7/17/98), Library of Congress (current as of 7/21/98).*

Table 2
Search Results for the Top 52 United States Newspapers by Circulation: Comparision of Burrelle's (B), Dow Jones Interactive (D), and NEXIS (N)

Newspaper	Date	B	D	N	Title
Arizona Republic	4/4/97		x	x	Digital TVs Expected To Sell Quickly, Despite $5,000 Price
Arizona Republic	4/3/97	x	x	x	Digital-TV era expected to begin within 2 years
Atlanta Journal and Constitution	4/3/97	x			Picture gets clearer: Digital TV in 2 years
Baltimore Sun	4/8/97		x	x	Digital TV channels called 'gift' to broadcasters; stations get $70 billion in free space on airwaves, ma
Baltimore Sun	4/3/97	x	x	x	FCC set to Okay timetable for digital TV
Buffalo News	4/8/97	x			Critics blast giving away digital channels
Buffalo News	4/3/97		x		Stations will be required to adopt digital TV
Chicago Tribune	4/3/97		x		Digital TV Now 2 Years Away; FCC Adopts Plan for New Television Technology
Cleveland Plain Dealer	4/3/97	x	x	x	FCC to OK digital TV plan
Dallas Morning News	4/4/97		x	x	Digital TV plan approved Network affiliates in D-FW must comply within 2 years
Dallas Morning News	4/3/97		x	x	FCC on verge of giving approval to plan for high-definition TV
Denver Post	4/9/97		x	x	Broadcasters may pay later for victorynow
Denver Post	4/8/97	x	x	x	Critics blast TV-airwaves 'giveaway'
Detroit News	4/4/97	x	x	x	Digital Television on the Horizon
Fort Worth Star Telegram	4/8/97		x	x	New free airwaves space raises question of payment for use
Fort Worth Star Telegram	4/4/97		x	x	Digital TV to debut in 2 years, FCC says The sets, which offer cinema-quality picture, will initially cc
Hartford Courant	4/4/97	x			FCC vote puts digital TV in sharper focus
Houston Chronicle	4/9/97	x	x	x	Smooth surfing could end; Broadcasters' wins may prove costly
Houston Chronicle	4/4/97		x	x	Converting to a digital revolution/Quality TV picture has hefty price
Miinneapolis Star Tribune	4/8/97	x	x	x	Giveaway or loan? Giving of channels to broadcasters for digital TV raises a storm or protest
Milwaukee Journal Sentinel	4/3/97	x	x	x	Digital TV mandate expected today
Ne w Orleans Times-Picayune	4/8/97	x	x	x	Free channels to broadcasters labeled 'giveaway'
Ne w Orleans Times-Picayune	4/3/97		x	x	Digital TV Plan Expected to get good reception today
Orange County Register	4/9/97	x	x	x	Digital-TV 'gift' gets uneven reception
Orange County Register	4/4/97	x	x	x	FCC approves plan for digital television
Orlando Sentinel Tribune	4/8/97		x	x	Critics: Broadcasters get deal in switch to digital but broadcasting companies say the government will
Orlando Sentinel Tribune	4/3/97		x	x	Broadcasters defend campaign contributions; A campaign fund raising watchdog group says broadcas
Philadelphia Inquirer	4/8/97	x			Digital--largesse or just a loan?
Philadelphia Inquirer	4/3/97	x			Study: Broadcasters shape policy
Philadelphia Inquirer	4/3/97	x	a	a	FCC to approve digital TV today
Portland Oregonian	4/4/97	x	x		FCC approves conversion of TV to digital
Rocky Mountain News	4/8/97	x	x	x	Critics view TV channels as giveaway
Rocky Mountain News	4/4/97	x	x	x	FCC Plugs in Digital TV
Rocky Mountain News	4/4/97		x		Government's Timetable
Rocky Mountain News	4/3/97		x	x	FCC to Mandate Digital TV by '99; Biggest Industry Advance Since Color
Sacramento Bee	4/4/97		x	x	Digital TV Era Clicks On
San Antonio Express-News	4/4/97		x		Pricey digital TV wins FCC's nod
San Diego Union-Tribune	4/8/97		x	x	Airwave allocation called big giveaway: Sale of TV spectrum could raise billions
Seattle Times	4/3/97	x			FCC wants digital TV to be ready in 2 years; biggest step since color
Seattle Times	4/3/97		x	x	FCC: Digital TV 2 Years Away; It's Biggest Step Since Color--You'll Need A Decoder for Your Old
St. Louis Post-Dispatch	4/8/97		x	x	Critics call TV plan huge giveaway
St. Louis Post-Dispatch	4/3/97		x		Networks' Political Influence Questioned
St. Petersburg Times	4/3/97		x	x	Reader's Digest Expands
		21	37	31	

43 articles cited the AP from 26 different newspapers
6 only by Burrelle's
6 by Dow Jones that not in Nexis
0 in Nexis that not in Dow Jones.
15 in online newspapers not found by Burrelle's

Table A
TV News Programs
(ABC, CBS, CNN, FOX, NBC, PBS)

		Verbatim Transcripts				Abstracts	
Network	News Program	Burrelle's	Dialog	Dow Jones[28]	NEXIS	Video Monitoring Service (VMS)[29]	Vanderbilt Archives
ABC	20/20	7/90...	1/97...	12/96...	1/90...	5/96...	
	Business World	7/90-3/93			1/90...	5/96...	
	Day One	3/93...				5/96...	
	Good Morning America	7/90...	1/97...	12/96...		5/96...	
	News Specials	7/90...				5/96...	
	Nightline	7/90...	1/97...	12/96...	1/90...	5/96...	9/88...
	Primetime Live	7/90...	1/97...	12/96...	8/89...	5/96...	
	This Week	7/90...	1/97...	12/96...	1/90...	5/96...	
	Turning Point	7/93...	1/97...	12/96-10/97		5/96...	
	Weekend Report	7/90...			1/90...	5/96...	
	World News Saturday	7/90...	1/97...	12/96...	1/90...	5/96...	12/78...
	World News Sunday	7/90...	1/97...	12/96...	1/90...	5/96...	12/78...
	World News This Morning	7/90...	1/97...	12/96...		5/96...	
	World News Tonight	7/90...	1/97...	12/96...	1/90...	5/96...	8/68...
CBS	48 Hours	2/90...		1/94...		5/96...	
	60 Minutes	2/90...		3/94...		5/96...	
	Burrelle's	10/90-8/94		6/94...		5/96...	
	America Tonight	10/90-8/94		6/94...		5/96...	
	Evening News	2/90...		1/94...		5/96...	8/68...
	Eye-to-Eye	6/93-8/95		1/94...8/95		5/96...	
	Face the Nation	2/90...		1/94...		5/96...	
	Morning News	2/90...		1/94...		5/96...	
	News Documentaries	2/90...				5/96...	
	News Specials	2/90...		5/94...		5/96...	
	Osgood File	1/94...		2/93...		5/96...	
	Public Eye with Bryant	10/97...				5/96...	
	Reports	6/93...				5/96...	
	Saturday Morning	8/97...				5/96...	
	Sunday Morning	2/90...		1/94...		5/96...	
	Sunday Night News	2/90-9/97		1/94...		5/96...	12/78...
	This Morning	2/90...		1/94...		5/96...	
CNN	Both Sides with Jesse Jackson			1/92...		5/96...	
	Capital Gang				3/92...	5/96...	
	Crier & Co.				3/92...	5/96...	
	Crossfire				1/92...	5/96...	
	Evans & Novak				1/92...	5/96...	
	International Correspondents				4/92...	5/96...	
	Larry King Live				4/92...	5/96...	
	News				1/90...	5/96...	
	Newsmaker Saturday				1/92...	5/96...	
	Newsmaker Sunday				1/92...	5/96...	
	Reliable Sources				4/92...	5/96...	
	Specials				4/92...	5/96...	
	The Big Story				4/92...	5/96...	
	Week in Review				4/92...	5/96...	
Fox	Fox News Sunday	4/96-4/97			10/96...	5/96...	

		Verbatim Transcripts				Abstracts	
Network	News Program	Burrelle's	Dialog	Dow Jones[28]	NEXIS	Video Monitoring Service (VMS)[29]	Vanderbilt Archives
NBC	Brokaw Reports	11/89...				5/96...	
	Dateline	3/92...		6/94...	1/97...	5/96...	
	First Person with Maria	1/89...		6/94-9/95		5/96...	
	Meet the Press	11/89...		6/94...	1/97...	5/96...	
	News at Sunrise	11/89...		6/94...	1/97...	5/96...	
	News Documentaries	11/89...		1/94...		5/96...	
	News Magazine	5/94...		5/94-10/94		5/96...	
	Nightly News	11/89...		6/94...	1/97...	5/96...	8/68...
	Saturday Today	8/92...		6/94...		5/96...	
	Sunday Today	11/89...		7/94...		5/96...	
	Today	11/89...		6/94...	1/97...	5/96...	
PBS	Bill Moyers Journal	10/97...		10/93-10/94		5/96...	
	Frontline				1/90...	5/96...	
	Newshour with Jim Lehrer				1/82...	5/96...	
	Washington Week in Review	1/97...			1/90...	5/96...	

Source: Data compiled from Fultext Sources Online (Needham Heights, Massachusetts: BiblioData, January 1998), a telephone conversation with John Lynch, Director of the Vanderbilt Archives (July 30, 1998), and various sources at Video Monitoring Services (August 3, 1998).

---

[26] Based on Audit Bureau of Circulation figures for the six months ending September 30, 1997.

[27] *Part of Knight-Ridder Tribune Business News (source code KRTBN); may also be available in Publications Library as an abstract.*

[28] Dow Jones resells its transcript file to Westlaw, so the two databases have similar coverage

[29] Available only on Nexis. Some coverage goes back as far as January 1993. In fact, the Nexis documentation as of July 14, 1998 states that all VMS coverage starts on January 1993, but this is clearly not the case. For example, a search revealed that NPR's *All Things Considered* did not start till July 1993. VMS personnel said that comprehensive coverage of the titles on this list began on May 1996.