European Journal of Communication

http://ejc.sagepub.com

Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'Push Button' Content Analysis

David Deacon

European Journal of Communication 2007; 22; 5
DOI: 10.1177/0267323107073743

Published by:
SAGE
http://www.sagepublications.com

Yesterday's Papers and Today's Technology

Digital Newspaper Archives and 'Push Button' Content Analysis

■ *David Deacon*

ABSTRACT

■ This article considers the methodological implications of using digital newspaper archives for analysis of media content. The discussion identifies a range of validity and reliability concerns about this increasingly prevalent mode of analysis, which have been under-appreciated to date. Although these questions do not deny a role for the use of proxy data in media analysis, they do highlight the need for caution when researchers rely on text-based, digitalized archives. ■

Key Words content analysis, digital news archives, Lexis-Nexis, political communication, research methods

Introduction

> Who wants yesterday's papers? Nobody in the world (Mick Jagger and Keith Richards, 1967)

It is often claimed that news is a disposable commodity: conjured in a moment and rendered instantaneously irrelevant by the march of time and the unpredictability of events. However well known such an assertion may be, it is ill founded. Journalists draw heavily on a 'vocabulary of precedence' (Ericson et al., 1987) when integrating, managing and interpreting

David Deacon is Senior Lecturer in Communication and Media Studies in the Department of Social Sciences, Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK. [email: d.n.deacon@lboro.ac.uk]

European Journal of Communication Copyright © 2007 SAGE Publications (Los Angeles, London, New Delhi and Singapore) www.sagepublications.com, Vol 22(1): 5–25. [10.1177/0267323107073743]

contemporary occurrences. Galtung and Ruge (1965) once remarked that 'News is olds'. Although their comment mainly refers to the intuitive values and recollections that shape news professionals' routine practices, it also covers how journalists frequently resort to their clippings files (whether actual or virtual) when reporting an issue, institution or individual they have little familiarity with.

Beyond the newsroom, there are many others who share a keen interest in examining the historical traces of news coverage, in both the short and long term. Legions of pressure groups, politicians, public relations specialists and other issue entrepreneurs monitor how information is presented in the media arena, and news archives are a key research resource for academics across the humanities and social sciences, as a source of information, as a subject for investigation in their own right and as litmus of broader social, political and cultural trends.

There are three perennial issues concerning the archiving of yesterday's news. The first concerns *storage*. When news material is retained in its original format, logistical problems regarding the availability of space can become overwhelming. Other methods of manual storage, such as the use of micro-film or micro-fiche for printed material, can alleviate these difficulties to some extent, but even these require the dedication of a considerable amount of physical space (particularly when one considers the associated need for viewing and reprographic facilities).

The second issue concerns *information retrieval* – i.e. to what extent is it possible to locate specific pieces of information without resorting to indiscriminate and time-consuming manual trawls through general archive material? Some elite news archives have long provided facilities designed to avoid such a necessity. The most famous example in the UK is *The Times Index*, which was first published as the Palmer's Index in 1868. Recent research has identified its continuing value as a search engine for all *Times*-related publications, both on the basis of its considerable historical reach (the indices date back to 1796) and the thoroughness and detail of its content categorization (Pearson and Soothill, 2003). Nevertheless, indices of this quality are the exception rather than the rule, and even those that exist are only produced annually and therefore distributed many months after some of the material they reference was originally published. This impedes their utility for short-term information retrieval.

The third issue concerns *access*. Previously, anyone wishing to consult conventional news archives had to be physically present to examine material, with all the attendant inconvenience this can cause. In the UK, there has long been a paucity of comprehensive broadcast news and newspaper archives. For example, until recently, any researcher wishing to examine even recent coverage from the most popular newspapers in the UK had to depend on the resources of the British Library's newspaper collection at Colindale, North London, due to the lack of popular press holdings in other public and academic libraries across the country.

Innovations in computer and information technology offer ways of alleviating the problems associated with storing, retrieving and accessing news material. Newspaper and, to a lesser extent, broadcast content is now routinely stored in various digital formats, which means it can be searched comprehensively, quickly and (apparently) reliably, and in many cases can be accessed remotely by subscribers. Interest has subsequently grown in how these computerized search facilities might be used in the systematic content analysis of news coverage and there is an increasing number of studies that have based their investigations on electronic searches of these digital sources (e.g. Altheide and Michalowski, 1999; Grover and Soothill, 1999; Esser et al., 2001; Reid and Misener, 2001; Kerr and Moy, 2002; Cameron, 2003; Freudenburg et al., 1996; Domke, 2004). In many cases, this involves using a database for content identification – i.e. identifying and collating relevant news material on a chosen topic that is then subjected to further manual analysis – but there are other examples where search facilities have been used as the principal basis for more specifically analytical tasks. These include using the search engines to quantify the prevalence (or otherwise) of certain terms over time and analysing the ways key words may co-locate in news content.

This article raises methodological questions about this rise of digitally based, 'push button' content analysis. It is motivated in part by a concern that these matters have not yet been given sufficient attention in the embrace of this mode of analysis. Specifically, the article considers the strengths and weaknesses of the Lexis-Nexis online system, which is a US-based commercial service. Originally set up for law firms and financial sources, it has now become the media archive of choice for many academic and political sources across North America and Europe. Indeed, such is its market dominance in the US, it has gained a vicarious political significance in its own right. Recently described as 'a readily accessible institutional memory of what candidates and presidents have said and done' (Grimes, 2004: 5), many politicians have become conscious of the ways the resource can resurrect past words to haunt contemporary ambitions. In a *Washington Post* interview, the vice president Dick Cheney mentioned the service as a specific case for consideration in an increasingly competitive and complex multi-mediatized environment (*The Washington Post*, 2004). The assessment provided in this article is restricted to the Lexis-Nexis 'Professional' service offered to UK-based users.

'Things' not 'themes'

Any search of digitalized news archives has to be based on the use of key words. As with other databases, Lexis-Nexis permits Boolean searches to extend or restrict its range. This dependence on key words has methodological implications as it determines the kinds of content analyses that can be conducted.

In an earlier review of the use of Lexis-Nexis in media analysis, Soothill and Grover (1997) identified the related problems of generating 'false positives' and 'false negatives' through key word searches. 'False positives' refers to those occasions when a word has several meanings and a search identifies a number of spurious 'hits' in the list of items identified. One example given by Soothill and Grover is using the term 'rape' to investigate press reporting of sexual violence. This search would not only locate articles reporting this serious sexual offence, but also items referring to the plant 'rape' and to 'a division of Sussex as well as refuse in wine-making' (Soothill and Grover, 1997: 592–3). 'False negatives' refers to searches where the key-wording is too precise, thereby excluding significant amounts of relevant coverage. Here again, Soothill and Grover explain why a reliance on the word 'rape' would be inadequate for any longitudinal investigation of press reporting of sexual offences, as journalists avoided use of the term before the 1960s, preferring more oblique phrases, such as 'sexual defilement', 'serious sexual offence' and 'carnal knowledge' (Soothill and Grover, 1997: 593).

In Soothill and Grover's view, the problem of 'false negatives' is more serious than 'false positives', as the latter can be easily rectified by weeding out irrelevant articles. Nevertheless, they conclude that both errors 'can be diminished through careful piloting of the most effective search keywords' (Soothill and Grover, 1997: 592). In my judgement, the problem extends further than this. Put simply, key word searching is best suited for identifying tangible 'things' (i.e. people, places, events and policies) rather than 'themes' (i.e. more abstract, subtler and multifaceted concepts). Because of this, there are certain topics that may be readily analysed via manual content searches, but which can never be captured through exclusive dependence on key words. Furthermore, a failure to appreciate this limitation can potentially lead to erroneous conclusions. To illustrate these points, it is useful to provide an example from actual research I have conducted into UK news reporting of 'Quangos' (quasi-autonomous non-governmental organizations) and which combined computerized and manual searches of news content (see Deacon and Monk, 2000).

Quangos are public bodies that are appointed to office, rather than elected. In the UK, their numbers and responsibilities have increased exponentially over the last two decades, which has fuelled concern about their accountability. How might such a content analysis of news reporting of quasi-government be conducted through a key word search of digital archives? It would be technically possible to enter the name of every known quango into a search engine, but the logistical problems this would create would be so great as to obliterate any of the convenience that digital searches are supposed to deliver (more than 7000 organizations fell within the definition of quasi-government adopted in the research). One could conduct a search of coverage of selected agencies, but this pre-selection would mean that these examples could only be treated as illustrative rather than representative of the sector as a whole. An alternative strategy would be to use the key word 'quango' and map the frequencies and contexts with which the term is invoked across different news media and over time. This, indeed, was a preliminary task we undertook (see Deacon and Monk, 2000: 49–55) and the results showed that:

- Journalists used the term very rarely.
- When the term was applied to a particular public organization, the report almost invariably focused on some negative or controversial aspect of their operations.
- A similarly negative frame of reference was evident when the term 'quango' was used to address broader issues concerning quasi-government in general, e.g. emphasizing the lack of accountability of this mode of government, its inefficiency or secrecy.

From these findings one might conclude that journalists have little routine interest in either the specific actions or general principles of quasi-government, and that, when they do, they are deeply sceptical on both scores. But how valid are these conclusions?

A manual content analysis of mainstream news reporting of quasi-governmental bodies was also conducted alongside this computerized search. In this aspect of the study, any item that referred to any organization that could be technically defined as a quasi-governmental body was included, even if it was not referred to as such in the article. The results that emerged contrasted considerably with those from the key word search. First, the term 'quango' was rarely applied to describe quasi-governmental bodies (merely 1.5 percent of the organizations identified in coverage were labelled with this term). Second, quasi-governmental bodies attracted far more news coverage than other non-governmental organizations. Third, instead of being disparaged as feckless, corrupt or incompetent, these public bodies were more commonly presented as authoritative and dispassionate arbiters of public policy – engaged in public debates, but removed from the political fray. In the main, journalists seemed more interested in recording the public statements, decisions and interventions of these agencies than in interrogating their internal structures and operations.

This evidence reveals a contradiction in journalists' perceptions of, and engagement with, quasi-government in the UK, in which quangos are deemed suspect in principle, but reliable in practice. The salient point for this discussion is that this more nuanced understanding could not have been derived readily and convincingly through key word searching strategies. Indeed, the key word used here identified entirely atypical coverage.

Other commentators have raised similar concerns in relation to research on other topics. For example, Althaus (2003) claims that many critical analyses of media–state relations underestimate the extent of press autonomy because of their dependency on the 'proxy data' of Lexis-Nexis searches, rather than a comprehensive analysis of the entire population of news coverage. Robinson et al. (2005) echo a similar concern in their review of several recent studies of media coverage of the 'War on Terror' and the invasion of Iraq, which all relied on digital news archive searches. According to Robinson and his colleagues, this failure to engage with actual news coverage inhibited the development of 'a fully fledged frame analysis that might reveal a broader range of debate' and probably resulted in 'the under-measurement of press criticism' (Robinson et al., 2005: 956).

Linguistic not visual

A more evident limitation of text-based digital news archives such as Lexis-Nexis is the loss of the visual dimension of news. This is a significant omission as the size and positioning of text and the use of photographs and illustrations are key mechanisms by which news-makers dramatize reports, assist readers' comprehension, corroborate the 'truth' of a reported event and, sometimes, qualify, or even subvert, the linguistic substance of a related news item.

Linguistic and visual elements of news are closely linked, but should not be treated as identical. As Higgins (2003: 2), summarizing Kress and van Leeuwen (1996), states:

> Visual structures and linguistic structures both realise meanings. These in part overlap between the two modes but are also different; some things can be said only visually, others only verbally. The way in which meanings are realised will be different: language choices are between, for instance, word classes, tenses, and semantic structures; visual choices are between, for example, colours, camera angles, and compositional structures.

Commentators have remarked how media analysis has tended to privilege linguistic analysis over visual analysis (e.g. Cottle, 1998), and a reliance on digital archives can only reinforce this tendency and inhibit understanding of 'the ways that meanings in popular media texts are created through the inter-play between language and image' (Deacon et al., 1999: 195). This is particularly regrettable at a time when the visuality of news has gained in importance, through the more extensive use of colour photographs and illustrations, larger, dramatic headlines and other creative compositional techniques.

Texts not contexts

Digital key word searches identify lists of individual articles that contain any references to the phrases entered. This form of unitization fits neatly with the kind of thematic content analysis most commonly deployed in media analysis (Beardsworth, 1980), where an article is treated as the host for a range of factual, thematic and linguistic features that are subsequently quantified (Deacon et al., 1999: 118–19). But these texts do not exist in isolation. They often function inter-textually, and the context of their placement and relationship with other texts can tell us significant things.

A facetious illustration of this point is offered by a full-page apology published by the British *Daily Mirror* newspaper on 22 October 2002. This apology was made to an American businessman who is the biological father of a celebrity's child and who had been attacked by the *Mirror* for allegedly neglecting his paternal duties. In an unusually forthright and fulsome expression of contrition, the newspaper apologized for the 'mean spirited and inaccurate articles it had published' and for 'urging our readers to telephone Mr XXX, and to disturb him with derogatory remarks based on our inaccurate reports'. It continued:

> Our readers should know that Mr XXX is not the ignominious character that has been depicted by some in the media. He is a philanthropist and humanitarian who has dedicated himself to helping causes impacting children. . . . We at *The Mirror* wish to take responsibility for our inappropriate actions, and are pleased to have this opportunity to set the record straight. Once again, XXX, we're sorry. (*Daily Mirror*, 22 October 2002: 9)

As apologies go, it couldn't have been more abject. However, its sincerity was compromised by an article placed on the facing page with the headline 'Why Americans Can't Understand Irony or Sarcasm' (*Daily Mirror*, 22 October 2002: 8).

Recent events, not the distant past

The impetus for the creation of digital newspaper archives like Lexis-Nexis came from revolutionary changes in news production practices themselves. From the mid-1980s, the computerization of text inputting and advances in desk top publishing meant that full text computer files of the newspaper material could be saved and marketed on a commercial basis.

One implication of this is that the historical reach of most digital news archives is limited.[1] Table 1 itemizes the availability of past editions of individual UK national press titles on the Lexis-Nexis service. Only *The Independent*, *The Independent on Sunday The Times*, *The Sunday Times* and *The Guardian* provide content from the 1980s, and most titles only became available from 1998 onwards. Although the historical breadth of the archive is growing on a daily basis, as things stand it is an archive that covers the recent past rather than more distant events. While the constantly updated material makes the archive undeniably useful in monitoring contemporary events, it can be seen to reinforce what some have lamented as an ahistorical tendency in much contemporary media and cultural analysis (e.g. O'Malley, 2002).

A specific and related concern with the Lexis-Nexis Professional service is its failure to explain clearly the precise dates and details of its newspaper holdings. Information linked to the opening search screen states that its UK press coverage ranges 'from 2 January 1982 to current; varies by publication; see individual source descriptions for details'. However, to find the exact details for each title involves a convoluted analysis of the source directory.[2] The analytical implications of this obfuscation can be serious. For example, I recently had to correct a draft of a student dissertation that claimed to have identified a dramatic rise in the use of the term 'spin doctor' in the UK press coverage from the late 1990s. While the term has undoubtedly gained greater public currency over recent years, the exponential increase identified in this instance was mainly an artefact of the greater number of newspaper titles that became available after 1998.

Computer searches and the aura of infallibility: from validity to reliability

All the comments made thus far can be said to relate to questions of *research validity* – i.e. to what extent can key word based investigations of text-only databases adequately capture the subtleties and complexity of meaning making in the media? On their own, these considerations do not deny a role for this kind of analysis, they highlight the methodological implications and limitations of this mode of analysis. However, there is another set of

Table 1 Availability of UK national press titles in Lexis-Nexis

	Available from		*Available from*
The Guardian	14 July 1984	*The Observer*	7 October 1992
The Times	1 July 1985	*The Sunday Times*	1 July 1985
The Daily Telegraph	30 October 2000	*The Sunday Telegraph*	30 October 2000
The Independent	19 September 1988	*Ind. on Sunday*	19 September 1988
Daily Mail	1 January 1992	*The Mail on Sunday*	1 January 1992
Daily Express	2 October 1999	*Sunday Express*	2 October 1999
Daily Star	15 December 2000	*Daily Star Sunday*	15 September 2002
The Sun	1 January 2000	*News of the World*	26 July 1998
Daily Mirror	29 May 1995	*Sunday Mirror*	29 May 1995
		The People	2 Jan 1994

Dates accurate as of 25 April 2006.

questions that need to be considered when assessing digitally driven news analysis. These relate to issues of *research reliability* – i.e. the extent to which computerized searches produce consistent, reliable and replicable results over time. This matter has received little consideration, which may reflect the aura of infallibility that tends to be attributed to computer technology. Apart from their undoubted convenience, computerized search engines apparently remove human error from the research process, identifying each and any reference to a specific term no matter how peripherally located in a newspaper's pages or deeply buried in the substance of an article. But this should not be taken on trust. Human intervention is evident in the data entry phase, search engines may have varying levels of sophistication, and the comprehensiveness of the archives may be affected by complex issues associated with publishing rights and copyright.

Inter-archive reliability

A first step in assessing the reliability of digital news searches is to compare the results produced for an identical key word using different digital news archives. This approximates the sort of inter-coder reliability testing commonly deployed in conventional quantitative content analysis. Figures 1 and 2 compare the results of searches of Lexis-Nexis Professional and the Chadwyck Healey CD Rom newspaper archives using the key word 'quango' for *The Guardian* (from 1992 to 2001) and *The Times* newspapers (from 1996 to 2001).

The Guardian comparison shows a strong correlation between the annual search results for the Lexis-Nexis and CD Rom archive. The only notable discrepancy occurs in 1994, when the number of articles found through the CD Rom search exceeded those found for Lexis-Nexis by nearly 10 percent (337 items and 307 items, respectively) and in 1995, where the difference was around 8 percent (290 items and 268 items, respectively).

The results for *The Times* comparison, however, reveal greater disparity. The totals for the years 1996–1999 are close, but for 2000–1 the results differ considerably, with the Lexis-Nexis counts on this occasion exceeding those found for the CD Roms. In 2001 (the year with the greatest disparity in results), the Lexis-Nexis search identified 23 items that were not found with the CD Rom search, whereas the CD Rom search identified two items omitted from the Lexis-Nexis list. But if Lexis-Nexis outperformed the CD Rom in terms of identifying relevant material on this occasion, its list also contained some duplicated entries that inflated the count.

These discrepancies may seem inconsequential, but it should be appreciated that the key word used for the comparison here is rarely used

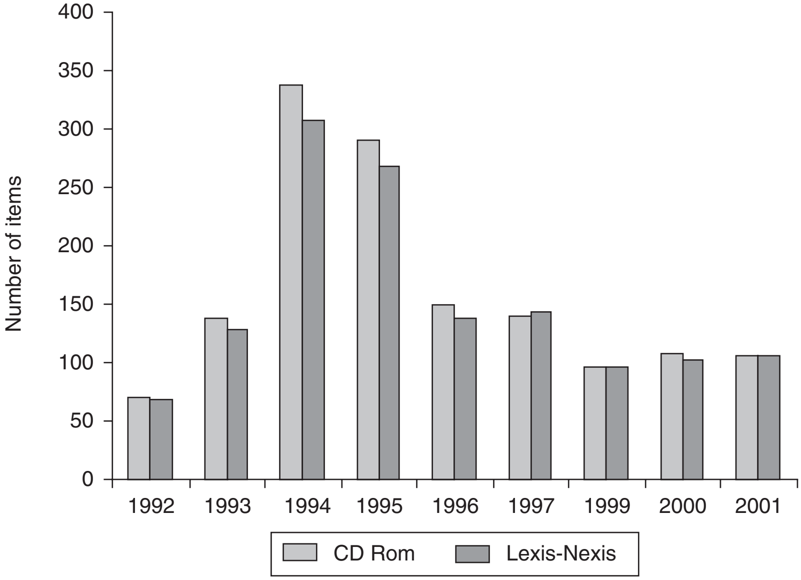

Figure 1 Comparison of the number of items identified referring to 'quango' in digital archives of *The Guardian* (by year)
Note: Data are missing for 1998.

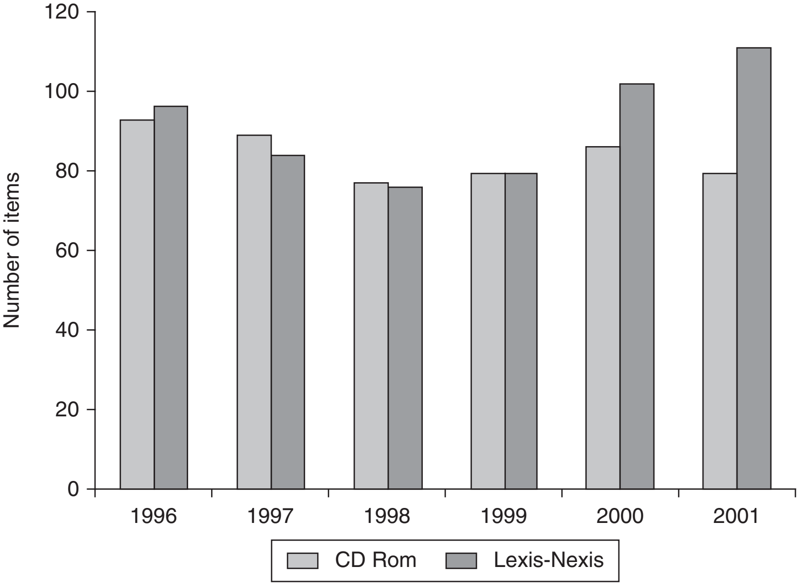

Figure 2 Comparison of the number of items identified referring to 'quango' in digital archives of *The Times* (by year)

in mainstream news coverage. Even greater discrepancies were found with key words more commonly used by journalists. For example, a search for articles referring to 'Tony Blair' for the 2001 Chadwyck Healey CD Rom edition of *The Times* identified 3239 items. An identical search on Lexis-Nexis identified 3410 items.

Intra-archive reliability

If this comparison suggests that Lexis-Nexis slightly outperforms its CD-based competitor in identifying content, its online format raises questions about its internal reliability. That is, to what extent do key word searches produce consistent results over time (intra-archive reliability)?

Unlike CD Rom based archives, researchers are licensed access to the online archive, they are not guaranteed access in perpetuity. A situation where an archive is 'loaned not owned' means there are no guarantees that (1) the content of accessible material will not be altered as a result of retrospective editorial actions (deletions, additions, modifications etc.), or (2) that the terms of permitted access will remain constant.

In terms of the first consideration, I found no evidence that subsequent editing of the Lexis-Nexis database produced inconsistent search results over time, as several key word searches I conducted in 2004 had identical outcomes in 2006. However, there are points of concern with respect to the second consideration. For example, in 2002, the Lexis-Nexis Professional service provided access to content from *The Daily Telegraph* from September 1988 onwards. In 2003, all of the paper's content published between September 1988 and 29 October 2000 was removed 'at the publisher's request'. Although referred to on the Lexis-Nexis site as a 'temporary' removal, this material is still absent three years after its removal. Also in 2003, the company introduced a new costing structure to its services, which meant that a range of non-UK newspaper titles and professional journals suddenly only became available via a higher premium service.

Changes of this kind have methodological implications, most obviously because they affect the scope of potential research. For example, the exclusion of foreign titles from the Professional service at a stroke removed opportunities for further cross-national media comparisons of the kind conducted by Esser et al. (2001) and Reid and Misener (2001). They also affect opportunities to reproduce earlier research findings. This is unfortunate as replicability is an important test of research reliability.

A further issue related to the intra-archive reliability of the Lexis-Nexis service is the consistency of results for identical key word searches conducted via different pathways offered by the search engine. In Lexis-Nexis, individual national UK titles can be searched by selecting them from either the 'UK

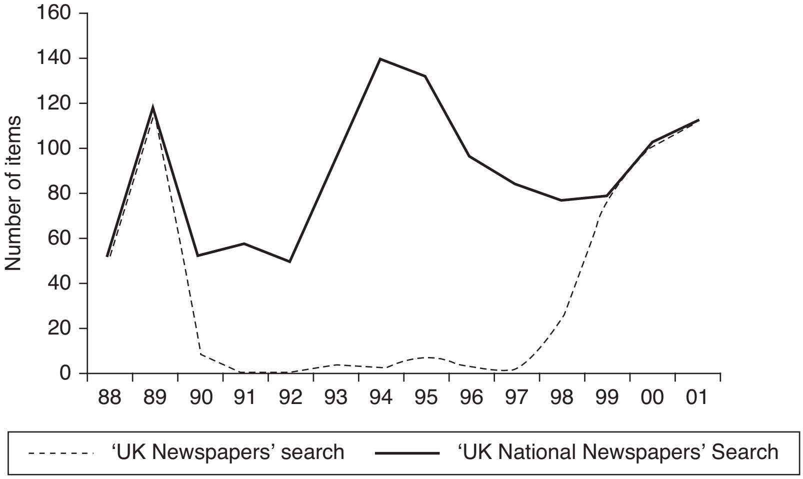

Figure 3 Comparison of the annual number of articles identified in *The Times* referring to 'quango' via searches of the 'UK Newspapers' and 'UK National Newspapers' pathways in Lexis-Nexis (1988–2001)

Newspapers' or 'UK National Newspapers' categories offered in the 'Sources' section of the opening search menu. In most cases, identical key word searches of titles by these different means produce consistent results, but I have found one striking discrepancy. Figure 3 compares the results of two searches conducted for the term 'quango' in *The Times* newspaper by these different pathways. The results for the search conducted via the 'UK National Newspapers' source option located a considerable number of articles containing the term during this period. However, when the same newspaper was searched via the 'UK Newspapers' option, hardly any articles referring to quangos were identified. Further key word searching suggested that the search engine is only partially accessing *The Times* archive via this pathway. For example, a search for items in the paper that contained the word 'government' (via the 'UK Newspapers' category) identified 3372 items for the three-month period 1 October–31 December 1998, and merely 1884 items for the 103-month period 1 February 1990 to 30 September 1998.

Double counts and no counts

It is common to find duplicated items in article lists produced by Lexis-Nexis searches. To give a dramatic example, a key word search using the term 'Tony Blair' of the content held for the *Daily Mail* between 1 January and 21 May 1996 generated a list in which every single article was duplicated. The reasons for double counts (and on occasions, multiple counts) are unclear. With some articles, it is indicated that a replicated item appeared in a later, or regional, edition of a national title, but this information is not consistently provided. Whatever the reasons for double counts, their potential presence means that raw quantification of coverage through searches can never be taken on face value. Just as one should check for 'false positives', so care must be taken to excise duplicated reports.

Although inconvenient, double counts can be easily identified and therefore do not pose a major reliability threat. Of greater concern is the potential for 'no counts', i.e. occasions where content was published but is not present in the Lexis-Nexis archive. These may represent isolated exclusions ('low level omissions') or more considerable absences ('high level omissions').

Low level omissions

The potential for some minor omissions is acknowledged on the Lexis-Nexis site, where it is stated that 'access to certain freelance articles and other features within this publication (e.g. photographs, classifieds, etc.) may not be available'. In my own experience, there have been occasions when I have searched the archive unsuccessfully for a particular item I know appeared in the published edition of a paper. For example, I was once unable to locate a controversial editorial that appeared in *The Mail on Sunday* (19 January 2001), flouting a German court injunction secured by the German chancellor prohibiting the paper from publishing details of his private life ('Sorry, Herr Schröder, But You Don't Rule Britain . . . At Least, Not Yet'). My initial assumption was that it had been excluded from Lexis-Nexis because of its questionable legality. However, having obtained a hard copy of the paper, I checked whether any other news, features or commentaries from that edition were missing from the Lexis-Nexis service. In a search restricted to the first 27 pages of the paper, I identified four other items that were absent. Although this may seem a small number, all of the missing items were substantial in size and collectively accounted for more than five pages of editorial copy.[3]

To assess whether this was an isolated case, I then selected three random days distributed five months apart and checked each item published in the hard copies of each of the UK national daily press to see whether it was present in the Lexis-Nexis archive.[4] Overall, 5 percent of items were found to be missing. Table 2 breaks this figure down by individual paper and sample day and also indicates what proportion these missing articles represented in terms of the total 'news space' of each edition. (Once again, the search was restricted to the major news and commentary sections of each paper and did not include readers' letters.)

Table 2 Missing items in Lexis-Nexis

	1 June 2005		*1 November 2005*		*1 April 2006*	
	Percentage of missing items	*Percentage of missing editorial space*	*Percentage of missing items*	*Percentage of missing editorial space*	*Percentage of missing items*	*Percentage of missing editorial space*
The Sun	7	4	2	1	2	1
Daily Mirror	–	–	11	7	3	12
Daily Star	5	3	3	3	–	–
Daily Express	4	2	4	2	–	–
Daily Mail	9	9	7	4	14	16
The Times	3	3	3	8	18	1
The Independent	2	2	2	0.5	–	–
The Guardian	7	6	–	–	–	–
The Daily Telegraph	Not available	Not available	Not available	Not available	21	7

Several points emerge from this comparison. Most papers had some material missing from the archive. In many cases, these absences were negligible, but with several titles they were considerable (e.g. *Daily Mirror*, 1 November 2005*; The Daily Telegraph* 1 April 2006; *Daily Mail*, 1 April 2006). On occasions, the two measures of missing coverage were not strongly correlated. For example, only 3 percent of items for the *Daily Mirror* published on 1 April 2006 were absent, but, due to their considerable size, these accounted for 12 percent of the total news space. In contrast, a high proportion of news items in *The Times* on 1 April 2006 were missing (18 percent), but, because these were very brief news items, they only accounted for 1 percent of the news space. Overall, there was no consistent pattern as to the comprehensiveness or otherwise of the records held for individual titles. For example, 6 percent of articles and 7 percent of news space were found to be missing from the archive for *The Guardian* on 1 July 2005, but everything was present for the two remaining sample days. In contrast, the *Daily Mirror* had no items missing for 1 July 2005, but significant amounts missing for 1 November 2005 and 1 April 2006. Finally, it was difficult to detect any consistency in the type of items missing from the database. To take the 1 April 2006 sample day as an illustration, missing items included news items ('Fizzy Drinks Pulled Off Shelves in Cancer Fear', *The Times*, 1 April 2006: 16), 'News in Brief' items ('Palestinian Factions Clash', *The Times*, 1 April 2006: 41), book serializations ('The Prince and the Funny Girl', *Daily Mail*, 1 April 2006: 50–3), celebrity exposés ('Dosh and Becks', *Daily Mirror,* 1 April 2006: 3) and general social commentaries ('Lost Age of Innocence', *Daily Mail*, 1 April 2006: 38–9).

The key point to consider is that, although these figures may seem small, once they are extrapolated over time, low level omissions can potentially accumulate into a considerable amount of excluded material.

A reassuring aspect of these findings is that no systematic pattern was evident in the omitted material. Therefore, it could be argued that low level omissions represent a type of random rather than constant sampling error; i.e. they have implications for the degrees of confidence we can have in any media sample we derive through these means, but they do not completely compromise its credibility. However, these tests do not completely rule out the possibility that there may be areas of the archive where exclusions are both patterned and considerable.

High level omissions

As a way of checking for larger gaps in the archive, I conducted random multiple key word searches of individual papers for discrete periods of time using very general terms ('said', 'today', 'Blair', 'sport' or 'government'). Given the sheer statistical improbability that any newspaper could print an edition in which none of its stories contained at least one of these ubiquitous terms, it was concluded that any search that produced a nil return indicated that no editorial content at all was available through Lexis-Nexis for that paper, for that period.

I must emphasize that this was an informal trawling exercise, as the logistics of systematically searching all titles for all periods were too formidable. Nevertheless, several random searches of titles and periods uncovered at least one gaping hole in the archive.

A key word search of the Lexis-Nexis holdings for the *Daily Mail* for the period 1 February 1996–30 May 1997 found 5136 items that made any reference to either 'said', 'today', 'Blair', 'sport' or 'government' in their content. Of these items, 2426 were duplications of other items identified by the search. (i.e. 47 percent of all items identified). But the most remarkable finding was that for 209 days (i.e. 54 percent of this 16-month period) *no items at all were identified via the key words*. For a further 81 days (i.e. 21 percent of this period) the search identified four or less items for an individual day (typically, a search using these key words identifies 200 plus items per day, per title).

It could be the case that this considerable lacuna is unique, but the fact that it was identified so quickly via a fairly unsystematic search does raise the possibility that there are other high level omissions in the service.

Unitization

A final reliability issue concerning Lexis-Nexis emerged unexpectedly through the process of assessing the extent of low level omissions. This concerned inconsistencies in the 'unitization' of material in the archive.

Unitization refers to the process by which one divides up a collection of material for subsequent analysis. As mentioned earlier, Lexis-Nexis stores its content in units that correspond closely to the kind of unitization commonly encountered in thematic content analysis. However, detailed comparison of the printed texts with their digital counterparts found inconsistencies in the unitization process. For example, on 1 April 2006 *The Times* published a news item about private funding of political parties. It had a major headline and text ('Tories Pay Back £5m to Hide Names of Lenders') and a related but distinct subsection with its own subheadline ('The 13 Backers Who Lent £16 million'). In this instance, both items were combined in Lexis-Nexis as one item. The coverage of the same story in *The Guardian* also contained a main and secondary item (Headline: 'A Farmer, a

Socialite and a Tycoon, but Who Are the Secret Names?', p.6; subheadline: 'The Lenders', pp. 6–7). On this occasion, however, the items were entered as separate items.

This kind of inconsistency was particularly evident in the treatment of columnists' work. In some cases, discrete topics discussed by the columnists were entered as separate items in their own right (e.g. Simon Heffer's column in *The Daily Telegraph* on 1 April 2006 was saved in Lexis-Nexis as seven distinct items). In other cases, they were segued into one meta-item (e.g. Simon Hoggart's equivalent column in *The Guardian* on 1 April 2006, which also discussed seven separate topics).

This inconsistency in the unitization of news content is worrying because it affects the statistical count produced by any key word searches and is far less easy to detect than doubly or multiply entered material.

Concluding remarks

The development and greater availability of digital news archives have resulted in a growing number of studies that base their media analyses on proxy data derived from these sources. These archives seem to offer the opportunity to quantify a large corpus of news material quickly, remotely and systematically; providing in seconds what would have previously taken months of perusing newspaper stacks or microfilm rolls.

However, there are methodological implications to this mode of analysis that have been insufficiently appreciated to date. These can be broadly differentiated as questions of research *validity* ('the integrity of conclusions derived from research' [Bryman, 2001: 30]) and *reliability* ('the extent to which results are consistent over time and an accurate representation of the total population under study' [Joppe, quoted in Golafshani, 2003: 597]). With regard to the former, four validity implications were discussed in this article: the difficulties of capturing complex thematic issues via key words; the problems of addressing the context of news content; the loss of the visual dimensions of news; and the reality that dependence on digital archives limits the historical reach of news analysis.

These matters apply to all text-based digital news archives. With regard to reliability considerations, this article focused on the performance of Lexis-Nexis, which is the most widely used digital news archive in social scientific research. A range of reliability concerns about the internal and comparative performance of this electronic archive were identified. These included inter-archive inconsistencies, intra-archive inconsistencies, multiply entered data, missing data and inconsistent unitization. In raising these matters, I do not mean to deny the considerable value of the Lexis-Nexis service as an information resource. But, by employing the service in quantitative content analysis, one is adapting its original purpose and thereby introducing a new range of stringent methodological criteria that need to be borne in mind when assessing its fitness for purpose.

'The elephant in the living room' is an English idiom used to describe the presence of a major issue that people would prefer not to acknowledge openly. The 'elephant' in this case is whether these validity and reliability concerns are so great as to deny any role for digital archives in the systematic quantitative analysis of news content. In my view, these results highlight the need for caution but do not preclude their use absolutely. There are a range of measures that can be used to increase the reliability of any analysis based on digital searches, such as checking for 'false positives' and duplicated items, scanning the titles and periods sampled for any high level omissions in data, and checking items for inconsistent unitization. Of course, such work takes time and care, thereby reducing the labour-saving benefits of this mode of analysis. (The one certain implication from these findings is that simple raw counts of coverage derived from key word searches must never be taken on face value.) Furthermore, these actions do not remove the possibility that there is some further sampling error, due to low level omissions in the database. Nevertheless, provided these issues are appreciated, and any subsequent evidential claims are modified on their basis, a role for 'push button' content analysis is still defensible.

However, it is vital to appreciate that a price is paid when media analyses depend heavily, or exclusively, on digital text. The evidence under analysis is proxy data and a lot of important evidence is lost in translation. For this reason, we should still aspire to analyse media content in its original form wherever possible, and where this is not possible, avoid casting necessity as a virtue.

Notes

This research was part of a broader ESRC-funded research project: 'Assessment and Development of New Methods for the Analysis of Media Content' (reference H333250014).

1. A notable exception to this is the Thompson-Gale *Times Digital Archive.* This contains digitalized facsimiles of every page 'as published' between 1785 and 1985. Aside from reproducing the visual dimensions of coverage, all text can be searched using key words.
2. (1) At the 'Search' screen select the 'power search option', then (2) select 'Browse source directory', (3) select 'news', (4) select 'individual publication', (5) select the alphabetical category for the title you are investigating, (6) click the 'i' icon alongside the individual title listed.

3. The five missing items were: (1) 'Stone Me! Look Who's Telling His Daughter's Boyfriend that He's Too Old for Her at 44 . . . Mick the Old Strolling Bone Himself' (a full-page celebrity news item on p. 7); (2) 'Germany's Chancellor in Court Bid to Gag MoS' (a full-page news item on p. 4); (3) 'Sorry, Herr Schröder, But You Don't Rule Britain . . . At Least, Not Yet' (a full-page leader editorial on p. 5); (4) 'Revealed: The Report that Left Tony Martin in Jail' (a one-and-a-half-page news item on pp. 22–3), and (5) '£25,000 bribe "Made Heath PM": EXCLUSIVE: Death-Bed Confession Reveals How the Tories Bought Harold Wilson's Election Plan' (a one-and-a-quarter-page news story, pp. 12–13).
4. My thanks to Ben Oldfield for his assistance with this task.

References

Althaus, S.L. (2003) 'When News Norms Collide, Follow the Lead: New Evidence for Press Independence', *Political Communication* 20(3): 381–414.

Altheide, D. and R.S. Michalowski (1999) 'Fear in the News: A Discourse of Control', *Sociological Quarterly* 40(3): 475–503.

Beardsworth, A. (1980) 'Analyzing Press Content: Some Technical and Methodological Issues', pp. 371–95 in H. Christian (ed.) *Sociology of Journalism and the Press*. Keele: Keele University Press.

Bryman, A. (2001) *Social Research Methods*. Oxford: Oxford University Press.

Cameron, P. (2003) 'Molestations by Homosexual Foster Parents: Newspaper Accounts versus Official Records', *Psychological Reports* 93(3): 793–802.

Cottle, S. (1998) 'Analyzing Visuals: Still and Moving Images', pp. 189–224 in A. Hansen, S. Cottle, R. Negrine and C. Newbold (eds) *Mass Communication Research Methods*. Basingstoke: Macmillan.

Deacon, D. and W. Monk (2000) 'Executive Stressed: News Reporting of Quangos in Britain', *Harvard International Journal of Press/Politics* 5(3): 45–66.

Deacon, D., M. Pickering, P. Golding and G. Murdock (1999) *Researching Communications: A Practical Guide to Methods in Media and Cultural Analysis*. London: Arnold.

Domke, D. (2004) *God Willing? Political Fundamentalism in the White House, the War on Terror and the Echoing Press*. London: Pluto Press.

Ericson, R.V., P.M. Baranek and J.B.L. Chan (1987) *Visualizing Deviance: A Study of News Organizations*. Milton Keynes: Open University Press.

Esser, F., C. Reinemann and D.P. Fan (2001) 'Spin Doctors in the United States, Great Britain and Germany: Metacommunication about Media Manipulation', *Harvard International Journal of Press/Politics* 6(1): 16–45.

Freudenburg, W.R., C.L. Coleman, J. Gonzales and C. Helgeland (1996) 'Media Coverage of Hazard Events: Analyzing the Assumptions', *Risk Analysis* 16(1): 31–42.

Galtung, J. and M. Ruge (1965) 'The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers', *Journal of International Peace Research* 1: 64–91.

Golafshani, N. (2003) 'Understanding Reliability and Validity in Qualitative Research', *The Qualitative Report* 8(4): 597–607; at: www.nova.edu/sss/QR/QR8-4/golafshani.pdf (accessed 8 May 2006).

Grimes, C. (2004) *The Press and Presidential Politics*. Syracuse: Campbell Public Affairs Institute, Syracuse University, New York; at: www.campbellinstitute.org

Grover, C. and K. Soothill (1999) 'Bigamy: Neither Love nor Marriage, but a Threat to the Nation?', *The Sociological Review* 47(2): 332–44.

Higgins, J. (2003) 'Visual Images in the Press: The Case of the Okinawa G8 Summit', presented to 'Knowledge and Discourse: Speculating on Disciplinary Futures', 2nd International Conference, web proceedings, July; at: ec.hku.hk/kd2proc/proceedings

Kerr, P.A. and P. Moy (2002) 'Newspaper Coverage of Fundamentalist Christians', *Journalism and Mass Communication Quarterly* 79(1): 54–72.

Kress, G. and T. van Leeuwen (1996) *Reading Images: The Grammar of Visual Design.* London: Routledge.

O'Malley, T. (2002) 'Media History and Media Studies: Aspects of the Development of the Study of Media History in the UK 1945–2000', *Media History* 8(2): 155–73.

Pearson, J. and K. Soothill (2003) 'Using an Old Search Engine: The Value of The Times Index', *Sociology* 37(4): 781–90.

Reid, W.J. and E. Misener (2001) 'Social Work in the Press: A Cross National Study', *International Journal of Social Welfare* 10: 194–201.

Robinson, P., R. Brown, P. Goddard and K. Parry (2005) 'War and Media', *Media, Culture and Society* 27(6): 951–9.

Soothill, K. and C. Grover (1997) 'Research Note: A Note on the Computer Searches of Newspapers', *Sociology* 31(3): 591–6.

The Washington Post (2004) 'The Strong Silent Type', 18 January: DO1.