RESEARCH PAPER

Tobacco industry documents: comparing the Minnesota Depository and internet access

E D Balbach, R J Gasior, E M Barbeau

See end of article for authors' affiliations

Correspondence to: Dr Edith Balbach, Community Health Program, Tufts University, 112 Packard Avenue, Medford, Massachusetts, USA; edith.balbach@tufts.edu

keyword: tobacco document searching

Received 13 June 2001 and revision requested 24 August 2001. Accepted 29 September 2001

Tobacco Control 2002;**11**:68–72

Objective: To assess the comparability of searches conducted on two publicly available tobacco industry document collections: hard copies housed and maintained by a neutral party in the Minnesota Depository and electronic copies available through tobacco industry maintained websites.

Methods: We conducted a set of searches in Minnesota and then conducted the same searches using the industry websites. We matched documents by Bates number, weeded out duplicates, and coded documents that were unique to either collection as major, minor, or trivial.

Results: Among hundreds of documents produced by several searches, we found only four unique major documents in the Minnesota Depository. By contrast, we found 62 unique major documents using the websites.

Conclusion: These results suggest that researchers can rely on industry websites while waiting for improved access resulting from searching, indexing, and document storage administered by the tobacco control community. Searching the tobacco industry websites is at least as good as searching in Minnesota and may in some instances actually be better. Four smaller subcollections, however, can only be searched by hand in Minnesota.

With the public availability of close to 40 million pages of documents produced by the tobacco industry in response to litigation, public health research on the tobacco industry has changed dramatically. Researchers who once had to rely solely on observation of tobacco industry behaviour can now also have access to the written record of tobacco industry strategies. With the support of grants from a variety of funding agencies, including the National Cancer Institute and the American Cancer Society, researchers at a variety of institutions are actively plumbing the document databases to answer a variety of research questions.

Initially, documents were only available in hard copy at the Minnesota Depository, a warehouse in Minneapolis, Minnesota that is managed by a neutral party. Subsequently, according to the terms of the Master Settlement Agreement, increasing numbers of documents have become available through websites, many of which are controlled by the tobacco industry.

This paper addresses the question of whether researchers can rely exclusively on industry maintained websites to gather a meaningful set of documents without travelling to Minnesota to search the Depository.* This is a reasonable question given that the two sources of industry documents are maintained by different parties—the Minnesota Depository by a neutral party, the websites by the individual tobacco companies. The answer is important not only because of the costs involved in travelling to Minnesota, but also as a means of checking whether the tobacco companies' websites are not posting important documents.

We answered this larger question by breaking it into two parts. Firstly, would identical searches of the Minnesota Depository and the industry websites yield the same results? And secondly, of the unique documents discovered in one location or the other, how many were substantial to our research?

It should be noted that searches, both online and at the Depository, are time consuming and frustrating[1] and that researchers' reliance on web databases and the industry produced 4B index at the Minnesota Depository is only a stop-gap measure until the tobacco control community progresses further in its goal to manage and index the documents. This goal is facilitated by two major efforts. One is a recent grant from the American Legacy Foundation to the University of California at San Francisco (www.library.ucsf.edu/tobacco) to mount and maintain a documents database, and the other is Tobacco Documents Online, a site run by Michael Tacelovsky (www.tobaccodocuments.org) with the involvement of several recipients of grants from the National Cancer Institute. These efforts to create sites that are easier to use are essential, because under the terms of litigation settlements, the Minnesota Depository is scheduled to close in 2008, and the tobacco companies are no longer required to maintain their websites after 2010.

METHODS

Searching Minnesota and industry website databases

Each document in the databases is indexed with certain identifying information, which includes a unique identifier called a Bates number, and often—but not always—some information about the title, date written, author, etc. This information is stored in searchable fields. At the Minnesota Depository, the search engines are only capable of searching discrete fields. That is, if one wanted to see what document indexes contained the term "strategy", one must first select the field to be searched (for example, the title field) and then enter the desired term or phrase. Wildcard or truncated searches cannot be done at Minnesota, although there is a functional search feature that allows a researcher to find variations in the spelling of a term. For example, when typing in the word, "target", the index will show that the words

*It is theoretically possible to search the 4B index (used to index the Minnesota Depository) on the website maintained by the Centers for Disease Control and Prevention, and to order documents one finds there from the Depository without actually going to Minnesota. But this is a slow way to search the Depository because of the delay in seeing documents. Because of the high number of unimportant documents retrieved by any search, the ability to scan the documents immediately and copy limited numbers of them is important.

Table 1 Search strings as entered on the 4B index at the Minnesota Depository and websites by company

Blue collar title searches
American Tobacco: advanced search interface, title: blue and title: collar
Brown & Williamson: advanced search interface, title: blue and title: collar
Lorillard: title: blue and title: collar
Philip Morris: title: blue and title: collar
RJ Reynolds: TTL: blue and TTL: collar* (old search interface; interface changed in May 2001)
Tobacco Institute: title: blue and title: collar

Strategy (date limited between 1990-2000) title searches
American Tobacco: advanced search interface, title: strategy and docdate: between 19900101 and 19991231
Brown & Williamson: advanced search interface, title: strategy and docdate: between 19900101 and 19991231
Lorillard: title: strategy (date range searching unavailable, sorted documents by hand)
Philip Morris: title: strategy and [ddatev: 19900101-19991231]
RJ Reynolds: TTL: strategy and DDT>19900101
Tobacco Institute: title: strategy (date range searching unavailable, sorted documents by hand)

Tobacco Institute title and keyword searches
*Target**
title: target or title: targeted or title: targeting or title: targets or title: targeted
Tennis and Virginia Slims
title: tennis and title: slim or title: slims or title: slim' or title: vs
Blue Collar
title: blue and title: collar
Lifestyle
title: lifestyle or title: lifestyles
Strategy (date limited between 1990-2000) title searches
Minnesota: title: strategy and docdate: [19900101-19991231]
Web: title: strategy (date range capabilities not working; documents sorted by hand)
Strategy (no date limitation)
title: strategy

"targeting", "targets", "targeted", "targetted", and "targeting" are also present in records and thus worth searching.

The industry websites, on the other hand, can perform more advanced keyword searches that scan across all indexed fields. Therefore, if the term "target" was entered as an "all text field," "combined text field," or "keyword" search, all documents that contained the word "target" in, for example, the title, descriptor or filename fields would also be returned. In addition, the industry websites allow for truncated or wildcard searches, so "target*" or "target%" could be used to retrieve all of the permutations listed above.

For this study, the more comprehensive searches allowed by the industry websites presented a challenge, because we would not know how comparable the Depository documents were to those found through the website if we did not try to exactly replicate the searches. For this reason, we searched the Depository collection as thoroughly as possible and then replicated this search on the industry websites. If this search yielded more unique documents when it was done on the websites than when it was done on the Depository, then there was no need to compare the more sophisticated search on the websites to the best Depository search. We would know that the more sophisticated search would be better if the basic one already was.

Our basic study design, thus, was twofold. Firstly, we compared two title searches between the Minnesota Depository and industry website databases, including American Tobacco, Brown & Williamson, Lorillard, Philip Morris, RJ Reynolds, and the Tobacco Institute. Second, we wanted to explore more thoroughly one database—the Tobacco Institute—by running several more searches and comparing the results of title searches done on the Tobacco Institute website and on the Minnesota database. All searches were conducted between mid October and mid November 2000.

We chose sets of search strings that reflected the research aims of our investigation into tobacco industry targetting of persons of lower socioeconomic status by the tobacco industry. Searches on "blue collar" and "strategy" (limited to documents written between 1990 and 2000) were run across the six major industry databases. The search strings focusing on "targeting", "lifestyles", "Virginia Slims", and "strategy" (with and without limitations on the date written) were run within the Tobacco Institute databases, because we were interested in the politics of targeting across companies. We thought that the Tobacco Institute would be interesting on this point. A complete list of our search strings is contained in table 1.

Comparing lists

Our searches produced lists of documents identified by unique Bates numbers. We compared the lists of documents from the Minnesota searches with those from the industry website searches and eliminated documents with matching Bates numbers. Comparisons were done manually, with a team of two (RJG and a student worker) comparing the lists. The remaining documents thus had unique Bates numbers.

Knowing that multiple copies of the same document often exist within a database, we then weeded out documents that, although they had unique Bates numbers, were duplicates of a document common to both databases. We checked to be sure that these were exact duplicates by comparing first pages, checking for document length, and looking for marginalia. What appear to be similar documents can, in fact, be different because of handwritten notes.

Coding the documents

We received hardcopies of the unique Minnesota documents by ordering them from the Depository and downloaded hardcopies of the website documents. Most files for the websites were pdf files, with the exception of Brown and Williamson and American Tobacco, which were mif files. The importance of each unique document to our research project was then rated and assigned to one of the three coding categories by two members of the study team (EDB and RJG):

Table 2 Number of matches versus unique documents produced by title searches performed on the 4B index at the Minnesota Depository versus tobacco company websites.

Company	Documents with matched Bates numbers	Documents unique to Minnesota title swarch	Documents unique to website title search
American Tobacco	73	6	5
Brown & Williamson	138	44	22
Lorillard	91	20	90
Philip Morris	526	49	352
RJR	547	0	147
Tobacco Institute*	629	0	217
Total	2004	119	833

The data reflects the results from the "blue and collar" and "strategy" (date limited) searches as well as to the four additional searches performed on the Tobacco Industry databases.
*The data on the Tobacco Institute includes the results of four additional search strings: (1) "target or targeted or targeting or targets or targeted"; (2) "tennis and (slim or slims or vs)"; (3) "lifestyle or lifestyles"; and (4) "strategy" without date range limitations.

- Major: Reports, memos, letters, budgets, or other items whose content is substantial. That is, it contributed in a material way to our research on its own, without reference to other documents.
- Minor: While not trivial, the document does not appear to have the length or substance of items coded as "major". A minor document may have suggested a lead for future research but had little inherent value. For example, a document might have been used as a slide in a presentation of a report, but lacked the substance of the full report detailing strategy. By searching on various elements in the document, such as Bates number, author, or date, it was possible, however, to find a more important document.
- Trivial: (1) Documents produced by an organisation outside of the tobacco industry, such as newspaper stories, periodicals, congressional hearing reports, and pamphlets; (2) inside materials that carry no information, such as file folders, meaningless cover memos, and internal operations memos such as staff training notes, copies of receipts, and phone bills; (3) document fragments in which there was no real content and no leads on how to find the main document; and (4) handwritten notes that could not be deciphered regardless of study team effort (there were very few cases of such notes).

When in doubt, documents were coded into the higher category. For example, a document on the border between trivial and minor would be coded as "minor". The coders agreed on over 90% of the documents; in cases where they disagreed, they discussed the document to achieve consensus about its coding. While this coding scheme represents a "rough cut" at the sorting of the documents, it was a simple way to assess the relevance of the documents for our research project.

RESULTS

Table 2 presents the number of documents found through title searches on the websites and at the Depository, indicating the number of common and unique documents. Generally, more unique documents were found by title searches on the websites than at the Depository. The exception is American Tobacco, where the numbers were 6 and 5, Depository versus website, respectively.

Table 3 identifies how many of the documents found were duplicates and how many of the remaining documents were coded as "major", "minor", or "trivial". Of the four major unique documents found in Minnesota, two were from Lorillard and two were from Philip Morris. Even with the more intensive searching of the Tobacco Institute index at the Depository, no major, unique documents were found there. By contrast, major, unique documents were found at every website, except Brown and Williamson.

For the Tobacco Institute website, we did try searching by keyword to see how many more documents would be retrieved compared to the more restrictive title search. We found an additional 70 major unique documents — including three for "lifestyle", 20 for "targeting", three for "Virginia Slims tennis", and 44 for "strategy"— 69 minor documents, 182 trivial documents, and 198 duplicates. All searches were conducted using the same search strings, with the title field indicators omitted (table 1).

DISCUSSION

The results from our search strategy comparisons indicate that, when faced with a choice of travelling to Minnesota or searching the tobacco industry websites, researchers can safely choose to search the industry websites. While the

Table 3 Number of major, minor, and trivial unique documents produced by title searches performed on the 4B index at the Minnesota Depository versus tobacco company websites.

	Documents unique to Minnesota title search				Documents unique to website title search			
	Major	Minor	Trivial	Duplicate	Major	Minor	Trivial	Duplicate
American Tobacco	0	1	4	1	1	0	2	2
Brown & Williamson	0	4	4	36	0	1	10	11
Lorillard	2	4	3	11	10	16	16	48
Philip Morris	2	4	14	29	25	43	89	195
RJR	0	0	0	0	8	13	55	71
Tobacco Institute**	0	0	0	0	18	10	101	88
Total	4	13	25	77	62	83	273	415
	Grand total: 119				Grand total: 833			

The data reflects the results from the "blue and collar" and "strategy" (date limited) searches as well as to the four additional searches performed on the Tobacco Industry databases.
*The data on the Tobacco Institute includes the results of four additional search strings: (1) "target or targeted or targeting or targets or targeted"; (2) "tennis and (slim or slims or vs)"; (3) "lifestyle or lifestyles"; and (4) "strategy" without date range limitations.

Table 4 Tobacco industry document resources

Litigant parties	Internet addresses for document repositories*
Philip Morris, Incorporated Document Site	www.pmdocs.com
RJ Reynolds Company Online Litigation Document Website	www.rjrtdocs.com
Lorillard Tobacco Company Document Site	www.lorillarddocs.com
American Tobacco Company	www.bwdocs.aalatg.com
Brown & Williamson Litigation Discovery Website	www.bwdocs.aalatg.com
Tobacco Institute	www.tobaccoinstitute.com
Council for Tobacco Research	www.ctr-usa.org/ctr
Additional sources for document collections	**Internet addresses or phone number**
Tobacco Control Archives at University of California, San Francisco: Current Collections (500+): Brown & Williamson Collection: The Cigarette Papers Joe Camel Campaign: Mangini v. RJ Reynolds British-American Tobacco Company Collection	www.library.ucsf.edu/tobacco/
Tobacco Documents OnLine (TDO) Current Public Collections (500+ Documents): The Roswell Park Bliley Collection (Philip Morris and Tobacco Institute) The Roswell Youth and Marketing Collection Massachusetts Tobacco Control Program Documents USC Tobacco Industry Monitoring Project Collection	www.tobaccodocuments.org
The Centers for Disease Control, Public Access to Tobacco Industry Documents	www.cdc.gov/tobacco/industrydocs
The Minnesota Tobacco Document Depository Information Line	(800) 526-8886 (USA only)

*The website www.tobaccoresolution.com has links to these websites and serves as a useful "homepage" for your browser.

research community waits for improved access resulting from searching, indexing, and document management by the tobacco control community, searching the tobacco industry websites is at least as good as searching in Minnesota and may actually be better. Of the hundreds of documents compared, four unique major documents were found in Minnesota.

Although the content of the four documents was interesting, the effort to isolate and track them was time consuming and, in hindsight, probably not worth the effort. The comparison done in this research took the study team approximately 250 hours plus travel costs to Minnesota. Any search of the Depository will generally require travel to Minnesota, time to sort through the boxes, and photocopying charges. Web searches, by contrast, can be done from anywhere and provide instant looks at the documents, which can then be downloaded or printed.

The exception to this finding is the material contained in four sets of document collections housed in Minnesota that are inaccessible to the desk bound researcher. The Minnesota Depository houses two document collections produced for the Minnesota trial by Liggett and British American Company/Industry. Liggett settled out-of-court and, as a result, their 180 boxes of documents are exempted from the database. The roughly 30 boxes of British American Tobacco documents were also exempted from the database because, as a non-American company, they were covered by a different set of legal parameters.

The third collection solely accessible through the Minnesota Depository fall into one of the following three categories: (a) trial exhibits, (b) learned treatises (papers used to establish witness credibility), and (c) demonstrative exhibits (used to make a point—for example, a pair of freeze dried lungs).

Finally, there is a collection of recently "depriviliged" documents from the Minnesota trial. The "deprivileged collection" is a small subset of a larger "privilieged" document collection produced during the Minnesota trial. These documents are now stored in roughly 84 boxes produced by all settling defendants: 1 CTR, 2 Lorillard, 46 RJ Reynolds, 2 Tobacco Institute, 4 British American Company, 1 British American Tobacco (BAT) Co Industries, 11 Philip Morris, 5 American Tobacco, and 12 Brown and Williamson. Little is known about the contents of these boxes; the only known characteristic this subset of documents share is that they were once protected under attorney–client privilege. The only way to determine the contents of this collection would be to travel to Minnesota to manually sort through the boxes. Researchers can request the "Other jurisdictions" list by contacting the Depository to find out more about these collections.

In addition, three major collections are not represented either in Minnesota or on the tobacco industry websites. One of these is the BAT Collection in Guildford, UK, the second is the Bliley collection, and the third is the complete files of the now-defunct Tobacco Institute and the Council for Tobacco Research, which will eventually be sent to the New York State archive. Of this last collection, only the files of the Council for Tobacco Research are currently at the New York State Archive; a small number of Tobacco Institute files have been shipped. While the Bliley Collection is being indexed by the Roswell Park Cancer Institute and is available through Tobacco Documents Online, the other two collections are less available. Some Guildford documents are at the University of California, San Francisco website but researchers must travel to the UK to

What this paper adds

Two publicly available document collections hold more than 40 million pages of internal tobacco company documents made available as a result of litigation. Hard copies are housed in the Minnesota Depository and maintained by a neutral party; electronic copies are available through tobacco industry maintained websites. This paper addresses the question of whether a researcher can rely exclusively on industry websites to search for documents related to a research question or whether one must also search the Minnesota Depository to avoid missing key documents. Among hundreds of documents produced by several searches, we found only four unique major documents in the Minnesota Depository, all of which existed in fragmentary form on the industry databases. These findings suggest that researchers can rely on industry websites.

see most of the collection. A list of all major collections is contained in table 4.

A secondary question addressed by our research is whether searching industry websites by keyword, rather than by title, yields additional documents of value. Although keyword searching pulls up many trivial and duplicate documents, it also produces a larger quantity of major documents in our test searches. We believe that this gain is worth the extra weeding required. In addition, every website contains instructions on how to do advanced searching on that site. To maximise one's searching effectiveness, it is worth spending some time learning about each site's capacity.

We would like to offer one caveat in interpreting our study results. The results of this study are dependent on the topic areas searched and the dates on which the search was performed. We have no reason to think that our topics are either more or less likely to result in good website searches when compared to Minnesota Depository searches. It is, however, possible that there are areas in which searching at Minnesota might be more fruitful in yielding unique documents not found on the industry websites. For example, some materials, including oversized items or media, are only available in Minnesota.

The relative richness of the websites compared to Minnesota does NOT mean that searching them is easy. As anecdotally corroborated by other document researchers, we experienced the daily fluctuations in the availability and quality of the industry websites. This variability warrants use of systematic searching techniques and a working understanding of the unique features of each industry database.

The websites and the Minnesota Depository rely upon use of often incomplete indexes created by the tobacco industry. The difficulties involved in searching these databases has been documented elsewhere.[1] Two additional difficulties we encountered are worth mentioning. Firstly, searches are sometimes case sensitive in ways that differ from the text which appears on the screen. A last name that appears on the screen in all capital letters may, in fact, really be comprised of a mixture of lower and upper-case letters. Secondly, dates are not always entered correctly, with some documents having no dates at all; further, during the fall (autumn) of 2000 when the searches were performed, date range searching capabilities on some of the websites were often not available or did not consistently produce results. However, given these irregularities, we believe that the search engines available through the industry websites will still yield a relatively complete data set when patient, thorough, and systematic searching techniques are used.

ACKNOWLEDGMENTS

This research is funded by American Cancer Society grant no. TURPG-00-287-01-PBP. The authors wish to thank Valerie Rock, Emily DaSilva, and Mia Baron for their assistance on this project.

REFERENCE

1 **Malone RE**, Balbach ED. Tobacco industry documents: treasure trove or quagmire? *Tobacco Control* 2000;**9**:334–8.