SPECIAL COMMUNICATION

Tobacco industry documents: treasure trove or quagmire?

Ruth E Malone, Edith D Balbach

Institute for Health Policy Studies, School of Medicine, Department of Physiological Nursing, School of Nursing, Department of Clinical Pharmacy, School of Pharmacy, University of California, San Francisco, California, USA
R E Malone

Community Health Program, Tufts University, Boston, Massachusetts, USA
E D Balbach

Correspondence to:
Ruth E Malone, PhD, RN,
Assistant Professor of
Nursing and Health Policy,
Institute for Health Policy
Studies, Box 0936, Laurel
Heights Campus, University
of California San Francisco,
San Francisco, CA 94143,
USA; rmalone@itsa.ucsf.edu

Received 25 November 1999
and in revised form
10 May 2000
Accepted 16 May 2000

The release of over 27 million pages of internal tobacco industry documents as a result of discovery processes in The State of Minnesota and Blue Cross and Blue Shield of Minnesota versus Philip Morris *et al* and other legal cases has provided tobacco control researchers and advocates with unprecedented opportunities to understand more about the inner workings of the industry. Documents are available for public viewing at the Minnesota Tobacco Document Depository, which opened in Minneapolis in 1998, at the Guildford Document Depository in Guildford, England, and on the world wide web, accessible through http://www.TobaccoArchives.com/ and other sites. In addition, through websites, users can get access to documents produced under the state litigation in Washington, Mississippi, Florida, and Texas, and selections from the British American Tobacco documents housed at Guildford, UK. (see "other tobacco documents resources", below).

Though the vast majority of documents are from the Minnesota case, which resulted in the release of documents from Philip Morris, RJ Reynolds, American Tobacco, Lorillard, the Tobacco Institute, Brown and Williamson, and the Council on Tobacco Research, collections continue to become available in conjunction with other legal cases. Given the enormous numbers of documents that are available, the collections may prove to be either a treasure trove of information valuable for tobacco control research and advocacy, or a quagmire of quantity into which researchers sink in despair. In this article, we discuss differences between searching for documents at the depository and on various online sites and suggest some practical strategies that may help researchers be more productive while using these collections.

Documents at the Minnesota depository

The depository is located in Minneapolis in a business park near the University of Minnesota. The number of computer terminals available for public searching is limited, so it is usually necessary to make reservations in advance of a visit to ensure that a terminal will be available for your use. The depository is open Monday through Friday from 8 am to 8 pm.

Going to Minnesota to search the collections offers the advantage of being able to hold in your hand and read clear copies of the documents, as opposed to the pictures of the documents available on the web. For very old pages or documents with handwritten notes, this is sometimes essential. The depository also holds oversized and multimedia materials, such as videotapes that are not available on websites, and a number of boxes of documents indexed only by Bates (unique identification) numbers. These documents have no computerised searchable index; therefore, one must sift through the boxes of papers to identify anything of relevance—mostly a matter of luck and a quick eye for screening. Descriptions of these collections not on the database are obtainable by contacting depository staff.

The majority of materials at the depository, however, are accessible using a computerised index. But the availability of this index does not mean that it is easy to obtain and examine copies of the documents. Researchers must first find a document by using the index, generate a list of documents to examine, then identify by number and company the box in which the document is stored, fill out and turn in a request form, wait for depository staff to retrieve the box from storage, sort through the densely packed pages to locate the document by Bates number, mark the place of the document in the box and, finally, fill out another form to order a copy. Copies are expensive and copying must be done by depository staff in order to ensure that documents do not disappear. This lengthy process ensures that documents are both in the boxes and in order by Bates number, but it also ensures that research will proceed slowly.

Further slowing research efforts is the process of finding relevant documents in the database. The problem is that the documents are not indexed using a controlled vocabulary. A "controlled vocabulary" is the list of the subject headings under which documents in a collection are indexed. The use of a controlled vocabulary standardises the terms that an indexer will use, limiting the number of terms and collecting relevant documents under a single term or phrase. For example, a skilled indexer would assign the term "California" to any document having to do with California tobacco policy and politics regardless of the document's verbatim title; the tobacco industry generated index does not do this. In the industry generated index, if the document's title includes "Cal", "CA", "Calif" or "California", then this is how it appears. Searching for one of these abbreviations alone means that the other documents are missed. Similarly, a document could deal with "Proposition 99" (a major California tobacco control initiative) and not mention "California" at all in the verbatim title, in which case it would not be found by searching for California documents under any of these variations on the word "California".

At the depository, using the industry supplied index, known as the "4B" index (also available as a CD-ROM), the documents for each company can be searched by author, date, the verbatim title, recipient, persons copied, type of document, and Bates number. Each tobacco company, however, did its own indexing and there is some variation in the indexing. In addition, each company's file must be searched separately. That is, one cannot simultaneously search the RJ Reynolds and Philip Morris databases. In addition, the number of index terms that can be searched simultaneously is limited, although it is possible to search simultaneously in multiple fields (for example, document date, document type, document title), and to combine terms using Boolean operators (and, or, not). One great advantage for those with some patience and curiosity is that searching at the depository allows the researcher to scroll down and peruse the actual indexing terms in each category, something not available online. This is a significant feature, as we discuss below.

VERBATIM VEXATIONS

The use of verbatim titles for indexing means that misspelled words are used as indexing terms. For example, a Philip Morris index search for the word "cigarette" at the depository will retrieve a mere 70 284 documents, small potatoes in such an enormous database. However, the 4B index contains over 125 other index words which will retrieve other documents potentially relevant to this search, including *cigarettes* (a separate 32 595 documents not retrieved in the search for "cigarette" singular), *cigarette]*, which retrieves an additional five documents, *cigarettess*, which retrieves 1 document, and *cigarets*, which retrieves another 821 documents. Also included are *ciagerette*, *cicaret*, *cisarette*, *citarettes*, and *cigtest*. Each of these indexing terms retrieves a separate, distinct set of documents. Even searching for something so seemingly straightforward as all documents identified as "memos" is not uncomplicated. Over thirty versions of this word appear in the index, including *moemorandum*, *mamorandum*, *memo*, *memorandum*, *memeo*, *memeorandum*, *memoradum*, *memorandom*, *memorandum, draft*, *memoranum*, *memroandum*, *meorandum*, *mo,memorandum*, and others.

Truncated or "wildcard" searches (for example, searching for ciga* or mem*) are possible online (but not at the depository), and would presumably retrieve most of the documents. However, even with truncated searches, there is no way to know how many documents one may have missed because of misspelled indexing without perusing the actual list of index terms (possible at the depository, but generally not available online). For example, a search online for ciga* would miss more than 20 of the 125 index terms we identified by visually searching the index listings. At the depository, however, all these terms are reviewable.

A good indexer creates a standardised set of subject headings and assigns them to the document based on an understanding of the content of the document. To date, tobacco control advocates have been unsuccessful in obtaining access to the industry's own internal "4A" index (which is, no doubt, much more efficiently organised), so it is still necessary to use truncation and/or dig through the mind-boggling lists of terms.

Documents at the Guildford UK depository

Documents from British American Tobacco (BAT) and its subsidiaries, including Brown and Williamson, are housed at the Guildford Document Depository in the UK, operated by BAT. Arrangement to visit the depository must be made far in advance through the company's lawyers, and based on our experience and that of other researchers (see http://www.tobacco.org/Documents/981217guildford.html), is exceedingly difficult to arrange, even several months in advance. To date, we have not visited the depository and know nothing of its organisational procedures. Some Guildford documents are, however, available on websites as discussed below, and a searchable index to file titles is available at http://www.tobaccodocuments.org/guil/ and at http://www.tobaccodocuments.org/index.cfm .

Documents on web sites

Searching for documents via the web is efficient and convenient. Online searching allows more comprehensive searches, and is a good way to identify and review a set of documents relatively quickly. Terms of the Master Settlement Agreement of 1998 included a requirement that companies enhance the searchability of their documents indexes by adding fields if such information was available. This means that online indexes are often more sophisticated than that available at the depository itself.

While searching from one's own computer is, of course, less expensive than going to Minnesota, such searching does require a computer with substantial memory (or the capability to store documents externally) and retrieved document images may not be clear (the documents online are retrieved as scanned images). Downloading and printing documents found online can be very slow, not infrequently resulting in a variety of error messages. In addition, tobacco control advocates have reported that documents posted to industry maintained sites sometimes "disappear".

A number of sites offer access to the documents. The following brief review of several sites is not comprehensive, but provides a starting point for exploring online access to these collections.

One of the best places to start is the **Centers for Disease Control and Prevention** (CDC) website (http://www.cdc.gov/tobacco/industrydocs/index.htm) which offers a good summary of information about the documents, a glossary, access to the 4B index, and information about and links to industry documents sites (however, some of these universal record locators (URLs) may be outdated). The site also offers very clear and detailed instructions on searching, accessible by clicking the "help" button in the "basic" or "advanced" search screens. One important tip from this site: different companies have different ways of indexing Bates numbers, with some disregarding the space between two sets of numbers and others including it. The CDC site suggests that if a document is not retrieved in a Bates number search, one should try either deleting the space between the two sets of numbers or inserting a 0 into the space and searching again.

The 4B index available on this site is an advance over the index at the depository, in that the CDC has merged the indexes from all the tobacco companies, saving the effort involved in searching each one separately. However, it is not possible to search the list of actual index terms. The site also includes two collections that are text searchable, the Minnesota Select Set and a set of some 7000 documents (out of several million) from the Guildford, UK document depository. These collections have been indexed using Optical Character Recognition (OCR) software, so that it is possible to search anywhere in a document for words relevant to your research aims. However, the OCR software may not accurately capture everything; for example, it may miss or misinterpret handwritten characters.

Searching the 4B index on this site will not retrieve images of documents; it will only identify them and provide an index summary. However, it is possible to order scanned or copied documents online from the depository on this site by clicking on the link for "public copy request".

The main industry site at the time this paper was prepared was http://www.tobaccoarchives.com/; however, as the URL for this site has changed in the past, it may be necessary to use major search engines or to search other sites to find it if it changes again. This site includes links to all the industry documents sites. A useful feature of this site is a chart of available index fields for all companies, found at http://www.tobaccoarchives.com/infochart.html. The site also includes general information about the documents, the legal agreements, and a list of some of the cases under which they were produced.

Among industry sites, the **Philip Morris** documents site (http://www.pmdocs.com/) is probably the most easy to use. This site uses the Alta Vista search engine and offers 32 searchable fields, including persons mentioned and brands mentioned. No list of indexing terms is available, but it is possible to do compound searches, for example, entering a name in a compound search for a person will retrieve that person as author, recipient, copied, mentioned, attending, and noted in a document. It is also possible to do Boolean searches and truncated or wildcard searches (although on several recent visits the wildcard feature was not operative), and the site offers fairly detailed and clear instructions on searching procedures. The site offers both page by page and "view all" options, and for optimal printing, pdf versions of documents are available using Adobe Acrobat software, downloadable from the site.

The **RJ Reynolds** (RJR) site (http://www.rjrtdocs.com/rjrtdocs/frames.html) is slightly less user-friendly, but offers certain advantages. RJR offers simple, basic, and advanced search options in searching some 24 fields, including author, mentioned brand, and copyees. There is a limit of 20 search terms for a single field, but there is a combined field that will use a search string to search all the fields. Boolean operators can be used and wildcard searches are possible, using ? for single characters and * for multiple characters. On this site, it is important to read the field definitions in order to know precisely how terms must be entered; for example, the format for the field "copyees" is last name space first initial middle initial. A helpful feature of this site is its "terms lookup" feature, available for some fields, that allows scrolling through the list of index terms, much as is possible at the depository. However, one can only view 20 terms at a time, rather than being able to scroll through long lists as at the depository, which makes this feature more time consuming for generating possible search terms. The site offers an option to view all pages of a document, rather than having to click on each page separately, and also offers pdf files of documents.

The **Lorillard Tobacco Company** site (http://www.lorillarddocs.com/) is visually and operationally similar to the Philip Morris site, uses the Alta Vista search engine, and offers 30 searchable fields. Compound searches are possible and documents are available as pdf files.

Both **Brown and Williamson** and **American Tobaccco** are searchable from the same home page (http://www.bw.aalatg.com/). Each can be searched using either basic or advanced searching, using 27 fields that are similar to those available at the other sites. Netscape was unable to find a "plug-in" for viewing images at this site on a Macintosh computer; a Windows one worked, although the system gave frequent error messages and shut itself down. Boolean searches and wildcard searches (using the "%" sign) are possible, although it was necessary to use the advanced search in order to string together a series of possible title words. It is possible to print either all pages or a selected subset.

The **Council on Tobacco Research** (http://www.ctr-usa.org/) is different from the other sites in that it only allows one type of searching. It is, however, quite easy to use. On the form itself it indicates that "or" Boolean searches should be conducted by putting commas between search terms and "and" ones should be indicated by using an ampersand. There are 27 possible fields, including "grant number," which allows the user to trace a particular project. Wildcards are possible using "*".

The **Tobacco Institute** site (http://www.tobaccoinstitute.com/) was down for revisions at the time this paper was revised. We have not personally used this site recently, but in the past, it was similar to the other industry sites.

General search strategies

So how should one proceed in a document search? One approach might be to use the online 4B index (available at the CDC website) to see how large a pool of documents is retrieved across all company databases using various search strategies. This will give an idea of the size of the "population" of documents for a particular search term. If the number is too large and/or the search is retrieving large numbers of irrelevant documents, search terms can then be refined using search strings. Once search terms have been settled upon, it is necessary to go to each company's document site (or to the depository) to retrieve and view the documents.

In searching, plan to search broadly for terms that might have been used in a verbatim title. This requires thinking like a historian: that is, focusing not only on terms in current use, but also on terms that might have been used at an earlier point in time. Once a relevant document is retrieved, read it and consider other terms in the document that might have been used in a verbatim title in another document. The collections that have already been indexed by tobacco control advocates are excellent sources for relevant documents to use in this process. The process of finding documents through their verbatim titles is a slow one, only intermittently rewarded with finding interesting material.

This process is, however, extremely fast when compared to the next important searching step. Not all documents have a verbatim title assigned. This means that some important documents can *only* be found by searching under the author's name, or online in a "person" field or fields. For example, if Ellen Merlo, a vice president of Philip Morris, was involved in planning strategy around a piece of legislation in a particular state, then it may be necessary to search under her name as author and recipient around the relevant date. Many documents will be listed, including quite a few with no title at all. A thorough search means looking at each of these to learn whether any is relevant. This process takes even more time than searching for verbatim title words but can be a key to finding interesting material.

Using Boolean operators (for example, smoking AND health) can help narrow searches when, as is common, an unmanageable number of documents is retrieved in a broad search. However, it is important to be careful in constructing searches, particularly complex strings of searches in several fields, as the inadvertent placing of an "or" where an "and" was intended will make an enormous difference in the results. It can be useful to proceed gradually, adding new fields and/or search terms stepwise to successive search strings in order to fine tune the search. Finally, in searching the documents, it is essential to record one's search strategy in detail. Otherwise, it is easy to forget what has already been done in which database and to begin repeating the search.

Once documents are retrieved, an excellent thesaurus for indexing tobacco industry documents (developed by Americans for Nonsmokers' Rights) is located at the website of the University of California at San Francisco http://www.library.ucsf.edu/tobacco/. This thesaurus is used to index documents available at the website. A similar controlled vocabulary is used by Roswell Park Cancer Institute for documents at its website, accessible through http://www.tobaccodocuments.org/index.cfm. This website also provides access to other collections indexed by tobacco control advocates. The advantage of using collections indexed by tobacco control advocates is that they have been correctly indexed. (See "Other tobacco documents resources" below).

With the new document searching initiative funded by the National Cancer Institute, more tobacco control researchers will be searching the industry databases. Sharing effective strategies and making documents available to the public through correct indexing will increase the documents' usefulness for tobacco control efforts worldwide.

Other tobacco documents resources

- http://www.tobaccodocuments.org/index.cfm : Links to numerous document collections, including a fully searchable set of the House Commerce Committee (Bliley) documents, an index for the Guildford documents, and bookmarking capabilities for document research.
- http://my.tobaccodocuments.org : This site, which may soon merge with the site above (both developed by Michael Tacelosky), is a test site that combines several industry indexes (Philip Morris, CTR, TI, and Lorillard) into a single site. Other industry indexes are due to be incorporated soon. This site allows for searching for names and organisations, in addition to searching for documents. It also includes tools for augmenting documents with personal notes, combining them into timelines, and adding information about people and organisations.
- http://www.library.ucsf.edu/tobacco/ : University of California, San Francisco Tobacco Control Archives: access to several document collections, indexing information, policy reports, and other tobacco control research and resources.
- http://www.tobaccopapers.org/ : British American Tobacco documents and documents from Guildford, UK depository.
- http://www.house.gov/commerce/TobaccoDocs/documents.html:Over 39 000 "secret" documents the industry sought to exclude from public view, released by the House Commerce Committee, but not very

useful for searching in their present form. There is some other information about the documents on this site, but searches are better accomplished at the tobaccodocuments.org site, described above.

- http://www.cctc.ca/ncth/guildford/:Approximately 10 000 selected documents from the Guildford depository, acquired by Health Canada.
- http://www.tobacco.org/Documents/secretdocuments.html : A collection of links and information about the documents.
- http://www.library.ucsf.edu/tobacco/searching. html: Detailed instructions for searching industry document web sites from the University of California, San Francisco library.
- The Minnesota Tobacco Document Depository: 1021 10th Avenue SE, Minneapolis, MN 55414, USA. (800) 526-8886 (US only); MNDepo@aol.com
- Access to the Guildford depository: Martyn Gilbey, British America Tobacco, Globe House, 4 Temple Place, London WC2R 2PG, UK, telephone +44 20 7845 1466, fax +44 20 7845 2783; Erika Reid, telephone +44 20 7845 1460, fax +44 20 7845 2783.