Tobacco documents research methodology

Stacey J Anderson,[1,2] Phyra M McCandless,[2] Kim Klausner,[3] Rachel Taketa,[3] Valerie B Yerger[1,2]

[1]Department of Social and Behavioral Sciences, University of California, San Francisco (UCSF), San Francisco, California, USA
[2]Center for Tobacco Control Research and Education, University of California, San Francisco (UCSF), San Francisco, California, USA
[3]Library and Center for Knowledge Management, University of California, San Francisco (UCSF), San Francisco, California, USA

Correspondence to
Stacey J Anderson, Department of Social and Behavioral Sciences, Box 0612, University of California, San Francisco (UCSF), San Francisco, CA 94143-0612, USA;
stacey.anderson@ucsf.edu

Received 17 November 2010
Accepted 21 January 2011

ABSTRACT

Tobacco documents research has developed into a thriving academic enterprise since its inception in 1995. The technology supporting tobacco documents archiving, searching and retrieval has improved greatly since that time, and consequently tobacco documents researchers have considerably more access to resources than was the case when researchers had to travel to physical archives and/or electronically search poorly and incompletely indexed documents. The authors of the papers presented in this supplement all followed the same basic research methodology. Rather than leave the reader of the supplement to read the same discussion of methods in each individual paper, presented here is an overview of the methods all authors followed. In the individual articles that follow in this supplement, the authors present the additional methodological information specific to their topics. This brief discussion also highlights technological capabilities in the Legacy Tobacco Documents Library and updates methods for organising internal tobacco documents data and findings.

INTRODUCTION

As of January 2011, the Legacy Tobacco Documents Library (LTDL) contained more than 11 million documents (representing more than 60 million pages) created by major tobacco companies relating to their advertising, manufacturing, marketing, sales, political and public relations, and scientific research activities. These were internal industry documents that were made publicly available through litigation against the tobacco industry and are housed in the electronic library, the LTDL (http://legacy.library.ucsf.edu). Based upon questions initially posed to us by the US Food and Drug Administration's (FDA) newly-formed Center for Tobacco Products regarding menthol in cigarettes, we conducted analyses of these documents in order to assess the knowledge and research conducted by tobacco companies on menthol and its relation to the following: marketing and consumer perceptions; initiation; topography; dependence; potential health effects; and cessation.

This supplement's authors used the same basic research methodology in researching their papers. Here we present an overview of the methods all authors followed, and the authors then present the additional methodological information specific to their topics in the individual papers.[1–6]

Tobacco documents research began in 1995 with the publication of five articles in the *Journal of the American Medical Association*,[7–11] and the methodology for conducting tobacco documents research has developed since that time. The technology supporting tobacco documents archiving, searching, and retrieval has improved greatly, which gives tobacco documents researchers more resources than having to travel to physical archives and/or electronically search poorly and incompletely indexed documents.[12]

IMPROVING SEARCH CAPABILITIES: OPTICAL CHARACTER RECOGNITION (OCR)

Until OCR capability, which translates scanned images of handwritten or printed text into machine-encoded text, was added to the LTDL in 2004, only the document records, or metadata, were searchable. A metadata record contains the document's important descriptive information such as the title, author(s), dates and a variety of subject terms. LTDL metadata come from a variety of creators. The tobacco companies and their affiliated organisations (the Master Settlement Agreement (MSA) collections: American Tobacco, Brown & Williamson, Council for Tobacco Research (CTR), Lorillard, Philip Morris, RJ Reynolds and Tobacco Institute) create their own metadata which the University of California, San Francisco (UCSF) Library then collects from the industry websites via a 'spider' program (a software program that reads information on websites in order to create entries for a search index). LTDL staff correct the metadata occasionally when there are obvious errors, such as an incorrect document date, or when the metadata contain information of a personal nature, such as social security numbers or bank account numbers, which LTDL redacts when they are identified. In some cases, a third party vendor either created metadata based upon reading individual documents (the British American Tobacco (BAT) Collection) or applied metadata based upon information provided to UCSF by the courts or the tobacco company (eg, the Liggett & Myers, Canadian Tobacco Trials and other non-MSA collections). Metadata for the Research Collection, obtained from Tobacco Documents Online (TDO; http://tobaccodocuments.org), were included in the LTDL exactly as they appeared on TDO, with UCSF adding the 'date added to UCSF' and collection name information. Metadata for the Depositions and Trial Testimony Archive (DATTA) Collection were provided by the Center for Tobacco Use Prevention and Research in Michigan as collected from their variety of sources. One consequence of this patchwork nature of metadata creation is that the metadata were not developed uniformly with research questions in mind and therefore may not incorporate sometimes highly relevant information useful to addressing specific research questions.

Conversely, the full text of a document refers to the complete electronic text of a source, subject to

the limitations of OCR (eg, failure to capture handwritten marginalia or poorly printed documents accurately). With searchable full text using OCR software, keyword searches now allow researchers access not just to the metadata, but to the contents of the documents themselves. Combining the metadata and the OCR improves a researcher's ability to locate relevant documents and decreases the necessary time to do so. Although a helpful addition, OCR does not replace the metadata for identifying relevant documents, particularly the ability to search the metadata for specific fields, such as persons, organisations or products named.

BOOKBAGS: STORING, ORGANISING AND ANNOTATING RESULTS

The bookbag function allows researchers to save a document's descriptive information and download or email the records. The search results screen allows one to save document records and annotate any record with one's notes to a file called 'bookbag'. Records can be saved individually or an entire page of results can be sent to the bookbag all at once. The ability to annotate records for inclusion in the bookbag improves a researcher's ability to organise search results and the research memos derived from a specific set of documents. The bookbag feature also allows researchers to download and import search records directly into a citation management tool such as EndNote (http://endnote.com) or RefWorks (http://refworks.com), improving the ability to create accurate bibliographies.

METHODS EMPLOYED FOR THIS SUPPLEMENT

There are well established protocols and methods for searching documents based on the snowball technique in which previous searches inform subsequent searches.[12–15] The broad research questions one begins with are used to guide initial searches of the document collections. One generates a list of search terms likely to return documents relevant to the research questions, and from qualitative analysis of the contents of those documents, the researcher refines research questions and continues the iterative process of searching, analysing and refining as a coherent story emerges.

The researcher builds a relevant collection of documents by reading and analysing search results, and conducting 'snowball searches'[12] based on the contents of the documents returned in initial searches. The addition of full text searching capabilities based on OCR processing often supplements but does not replace the usefulness of searching specific metadata fields or keyword searching in the metadata by allowing us to find important individuals, organisations or products as well as words and phrases in the document text. Once relevant documents are located using the initial keyword searches, we do follow-up searches to locate additional similar documents using the names of key individuals, organisations and products listed in the metadata, document areas (physical locations in the companies' filing systems where the documents were stored), project names and dates, budgets, organisational charts and consecutive reference (Bates) numbers. For instance, a document may mention specific brand names that are important to understanding the research question. Or the author and/or recipients of the document may be involved in projects relevant to the research question. Gaining this additional information through qualitative analysis of the metadata and documents allows the researcher to refine the research question and fill out the analysis with more targeted documents data.

The qualitative methods employed in this research, which is used by historians and social scientists who study archival and documentary data,[16] involve iteratively reviewing data to construct an account that is as coherent, supported by the available documentary evidence and contextualised as possible within the limitations of the documents archives (see Limitations section, below). The manuscripts in this supplement are what Carter[17] referred to as 'A papers', defined as research papers that are primarily concerned with tobacco industry documents (as opposed to 'B papers', defined as not primarily reports of tobacco documents research). All of the papers in this supplement relied on our analyses of documents in the LTDL and on the published literature to identify how the current findings fit into the broader context of what is already known about the tobacco industry's interest in menthol. In two manuscripts,[1 2] the authors also examined various tobacco advertising archives (http://www.tobacco.org/ads; http://www.tobaccofreekids.org/adgallery; http://www.trinketsandtrash.org; http://lane.stanford.edu/tobacco/index.html) to compare final advertisements with the marketing planning documents.

It is useful to record one's search strategy and write research memos to aid in the analysis of several iterations of documents searches. Research memos containing direct quotes, search strategies and content summaries are used to build a picture of how the pieces fit together historically and conceptually. Supplemental searches are conducted throughout this process to answer increasingly focused questions. For example, the general question, 'How does the tobacco industry sell menthol cigarettes to young adults?' could be narrowed to 'How did Lorillard use focus group research to identify marketing messages likely to appeal to potential young adult female targets for the menthol market?'

Our initial search terms often yielded tens of thousands of results. The authors reviewed the first 50–350 documents returned based upon 'relevancy', defined as roughly the number of times a search term appears relative to the number of pages in the document. For example, a single page document with a search term appearing once would appear before a 10-page document with the search term appearing once. Although this definition of 'relevancy' is not foolproof and certainly will favour shorter documents over longer documents because of the term-to-page ratio, it provides an adequate picture of how relevant the results set will be to the search terms, particularly when there is a long list of search terms that narrow the results.

Various methods to reduce the possibility of missing an important group of documents in the returned results were employed. Beginning by screening the first 100 or so returned documents, those that have the most occurrences of the search term(s), tells the researchers if the combination of search terms is generally fruitful or if the terms return documents that are off topic. In some cases, returned documents were sorted by date (ascending and descending in subsequent sortings), and the first 100–200 were screened, followed by the next 100–200 in the next decade, and so on. This helped researchers gain an understanding of what had happened over time as well as identify patterns across time and which time periods showed the most activity on a specific topic.

In addition to these screening techniques, narrowing down the results from tens of thousands to our final sets is a function of eliminating duplicates and documents that were not useable for our purposes (such as articles and scholarly papers that the companies had copied and kept). It is common to return a large amount of duplicate documents that made up a percentage of the results, and weeding them out by fielded queries would

reduce the number of relevant returned documents. For example, a search on 'Jeltema' (surname of a PM employee) and 'menthol*' returned 2798 results. Within this set there were at least 647 duplicates, or 23% of the returned documents. Although some duplicate documents may contain important information in the form of handwritten marginalia, when scanning such large numbers of documents, one weighs the benefits of identifying the maximum number of relevant unique documents against the costs of missing a duplicate with marginalia that could also be relevant.

Finally, the researchers considered diminishing returns. If one finds 2 relevant documents in the first 50 results, 1 in the next 50 and none in the following 50, one often skips to a group of results further down the list of returned documents and scans for relevant documents in 50 results. If no relevant documents are found at this point either, one is likely to abandon that set of results and move on to the next combination of search terms. For example, for Yerger's article,[4] 48 queries were conducted, for which the entire results set of 32 of those queries were reviewed (67% of all queries). In an additional seven queries for this manuscript, a partial review such as those described above revealed no relevant documents. If scanning 200 results reveals no relevant documents, it is judged not very likely, although certainly possible, that relevant documents will be found.

Based on our initial screenings, documents that did not appear to be relevant to the research questions or duplicated documents already found were discarded. Relevance to the research questions was based on whether, upon electronically searching or reading a document, it included content related to the topic or our specific research questions. Documents that passed this screening and were determined to be worthy of further review were read and analysed, and through qualitative analysis and contextualisation of these, the researchers found specific themes emerging. Based upon those thematic findings, the researchers selected the documents that most accurately illustrated the general research findings to cite in the papers. Thus, it is the findings based upon analysis of the tobacco company's own statements that determines the selection of representative documents to be cited in the papers. Not cited were documents that summarised the thematic findings but not as eloquently, documents that supported the findings but were difficult to understand out of context, or documents that were deemed not relevant to the research questions after our further reviews.

LIMITATIONS

Qualitative documents research has some limitations. First, the sheer quantity of available documents (over 60 million pages) forces researchers to make decisions about which search terms retrieve the most relevant material, and establishing a comprehensive list of search terms capable of returning every document relevant to a topic is not possible. The LTDL is frequently updated as tobacco companies provide additional material and documents become available through litigation. Therefore, some relevant data in the archives will not have been included in the analyses.

Documents in the LTDL and similar tobacco documents archives are of unknown representativeness, due to the routine or ad hoc discarding of documents or purposeful document destruction on the part of the companies producing them. Therefore, individual documents or sets of documents necessary to establish the true context of a topic may be missing, increasing the risk that a topic may be only partially understood or misinterpreted. It is for this reason that multiple search queries are performed and analyses are conducted in an iterative and 'bootstrap' fashion, so as to gain as much context as possible within this limitation.

'Saturation' is the ideal situation in tobacco documents research. This refers to receiving the same documents resulting from many combinations of different search terms, signalling to the researcher that the most important documents have been found. In a short time frame, it is not always possible to achieve saturation, and it is always possible that even upon saturation there may yet be a set of relevant documents in the entire archive that have not been found.

Further, the metrics used by tobacco industry executives to measure a variable (for example, nicotine dependence or the success of a marketing campaign) are not constant across companies, studies, or time periods. There are necessarily holes in the record dependent upon the quality and quantity of data the companies recorded as well as what documents are released, or not released, to the archives.

Finally, there is evidence that the industry tried to hide its findings, although it is unclear from whom. For example, in a 1974 BAT memo about a visit to BIBRA, a toxicology consulting firm, it was noted that 'Reference to menthol should be omitted from such documents (invoices), which should refer generally to toxicity studies'.[18] Brown and Williamson used code terms, such as 'Kintolly', 'Tolkin', 'Harpat', 'Polar Bear', and 'Cenmap' when referring to menthol.[19] Acronyms were also commonly used, which are often unclear if the context is unknown.

Despite these limitations, tobacco documents research continues to make a unique contribution to public health's understanding of the tobacco epidemic—the knowledge, activities and intentions of tobacco industry insiders as stated in their own words.

Acknowledgements We thank Karen Butter, University Librarian, University of California San Francisco, for procuring the funding for this project.

Funding This research was supported by the US Department of Health and Human Services Contract HHSN261201000035I and by NCI grant no. CA113710-05.

Competing interests None.

Contributors All authors contributed to the design of the study and the drafting of the manuscript, and gave final approval to the manuscript.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Anderson SJ.** Marketing of menthol cigarettes and consumer perceptions: a review of tobacco industry documents. *Tobacco Control* 2011;**20**(Suppl 2):ii20—ii28.
2. **Klausner K.** Menthol cigarettes and smoking initiation: a tobacco industry perspective. *Tobacco Control* 2011;**20**(Suppl 2):ii12—ii19.
3. **Yerger VB,** McCandless PM. Menthol sensory qualities and smoking topography: a review of tobacco industry documents. *Tobacco Control* 2011;**20**(Suppl 2): ii37—ii43.
4. **Yerger VB.** Menthol's potential effects on nicotine dependence: a tobacco industry perspective. *Tobacco Control* 2011;**20**(Suppl 2):ii29—ii36.
5. **Salgado MV,** Glantz SA. Direct disease-inducing effects of menthol through the eyes of tobacco companies. *Tobacco Control* 2011;**20**(Suppl 2):ii44—ii48.
6. **Anderson SJ.** Menthol cigarettes and smoking cessation behavior: a review of tobacco industry documents. *Tobacco Control* 2011;**20**(Suppl 2):ii49—ii56.
7. **Barnes DE,** Hanauer P, Slade J, *et al*. Environmental tobacco smoke. The Brown and Williamson documents. *JAMA* 1995;**274**:248—53.
8. **Bero L,** Barnes DE, Hanauer P, *et al*. Lawyer control of the tobacco industry's external research program. The Brown and Williamson documents. *JAMA* 1995;**274**:241—7.
9. **Hanauer P,** Slade J, Barnes DE, *et al*. Lawyer control of internal scientific research to protect against products liability lawsuits. The Brown and Williamson documents. *JAMA* 1995;**274**:234—40.
10. **Slade J,** Bero LA, Hanauer P, *et al*. Nicotine and addiction. The Brown and Williamson documents. *JAMA* 1995;**274**:225—33.
11. **Glantz SA,** Barnes DE, Bero L, *et al*. Looking through a keyhole at the tobacco industry. The Brown and Williamson documents. *JAMA* 1995;**274**:219—24.
12. **Malone RE,** Balbach ED. Tobacco industry documents: treasure trove or quagmire? *Tob Control* 2000;**9**:334—8.

13. **Bero L.** Implications of the tobacco industry documents for public health and policy. *Annu Rev Public Health* 2003;**24**:267—88.
14. **Balbach ED,** Gasior RJ, Barbeau EM. Tobacco industry documents: comparing the minnesota depository and internet access. *Tob Control* 2002;**11**:68—72.
15. **Hirschhorn N.** *The Tobacco Industry Documents: What They Are, What They Tell Use, and How to Search Them: A Practical Manual*. Geneva, Switzerland: World Health Organization. http://www.who.int/tobacco/communications/TI_manual_content.pdf.
16. **Miles MB,** Huberman AM. *Qualitative Data Analysis: An Expanded Sourcebook*. 2nd edn. Thousand Oaks, CA: Sage Publications Inc., 1994.
17. **Carter S.** Tobacco document research reporting. *Tob Control* 2005;**14**:368—76.
18. **Binns R.** *Visit to BIBRA: 25th February 1974*. London: British American Tobacco, 1974. Bates No. 400990448/0449. http://legacy.library.ucsf.edu/tid/hjm10a99.
19. **Tinsley M.** *Brown & Williamson Tobacco Corp. Subjective Coding Project—Substance Glossary*. San Francisco, CA: UCSF B&W, 1989. Bates No. 1328.01. http://legacy.library.ucsf.edu/tid/qyc72d00.