MITCHELL | HAMLINE School of Law

William Mitchell Law Review

Volume 25 | Issue 2 Article 11

1999

Filtered or Unfiltered Information: Choices in How to Make the Minnesota Tobacco Document Depository Records More Accessible to the Public

Jeanne Weigum

Michael Ravnitzky

FILTERED OR UNFILTERED INFORMATION: CHOICES IN HOW TO MAKE THE MINNESOTA TOBACCO DOCUMENT DEPOSITORY RECORDS MORE ACCESSIBLE TO THE PUBLIC

Michael Ravnitzky†
Jeanne Weigum††

† J.D. May 1999 (expected), William Mitchell College of Law. Mr. Ravnitzky is a board member of the Association for Non-Smokers—Minnesota ("ANSR"). Mr. Ravnitzky has done extensive research into depository records since the first day the depository opened to the public.

†† Executive Director, ANSR. Ms. Weigum has led tobacco-control efforts in Minnesota for over 25 years. Ms. Weigum recently participated in the July 1998 Technical Assistance Meeting on Retrieving, Cataloguing and Analyzing Tobacco Industry Documents, sponsored by the U.S. Department of Health and Human Services, and is on the board of directors of Minnesota Partnership for Action Against Tobacco ("MPAAT").

I. INTRODUCTION

When the State of Minnesota and health insurer Blue Cross and Blue Shield of Minnesota engaged the tobacco industry in a much-publicized legal battle, both sides produced an arsenal of documentation so vast as to be nearly unfathomable. Finding several million pages of industry data unmanageable, the trial court ordered the creation of two document depositories.[1] The tobacco industry continues to be the centerpiece of worldwide legal, political, social, and medical debate. Consequently, the public's access to the millions of warehoused documents is imperative. Naturally, such access is highly dependent on an accurate, useful document retrieval system.

This Article describes the workings of the Minnesota tobacco depository and suggests necessary changes to ensure meaningful public access to the documents therein. Part II recounts the recent history of the depository and serves to underscore its emerging importance as a research site. Part III explains the composition of the Minnesota collection, detailing the content and describing methods of research. This section also identifies the depository's significant shortcomings, and addresses recent governmental policy objectives designed to improve access to the documents. Offering solutions, Part IV examines two categorizing strategies that aim to cure existing document retrieval problems. First discussed is an expert classification method, followed by a survey of document scanning techniques. A potential hybrid of the systems as a solution is also explored.

Finally, this Article proposes, in Part V, that the Minnesota tobacco depository's current indexing system is unsuitable for meaningful public access to vital industry data. Because current technology affords a number of viable solutions, the U.S. Department of Health and Human Services ("HHS"), tasked with making the documents available to the public, should undertake measures to expedite this goal.

1. *See* Consent Judgment, State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, 1998 WL 394336, at *3 (Minn. Dist. Ct. May 8, 1998); State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, slip op. at 2 (Minn. Dist. Ct. July 14, 1995).

II. BACKGROUND AND HISTORY

During the Minnesota tobacco litigation, the trial judge ordered the establishment of two document depositories: one warehoused in Minneapolis, Minnesota, and the other in Guildford, England.[2] Over thirty-three million pages of documents were provided by tobacco industry defendants to attorneys for the State of Minnesota and Blue Cross and Blue Shield of Minnesota; twenty-six million pages in the Minnesota depository and seven million pages in the Guildford depository. Thousands of hours of videotape, audiotape and hundreds of reels of microfilm were also provided during discovery. A painstakingly chosen subset of these documents, called the "Minnesota Select Set,"[3] formed the foundation for the Minnesota trial. Due to the exigencies of trial, only a tiny fraction of the Minnesota Select Set documents were actually introduced into evidence.[4]

---

2. The Guildford depository was established to allow the convenient storage of records from the British American Tobacco family of firms, usually referred to as BAT or BATCO. Under court orders, BAT is obligated to maintain the Guildford tobacco depository for ten years. *See* Consent Judgment, State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, 1998 WL 394336, at *3 (Minn. Dist. Ct. May 8, 1998). Recently, the trial court overseeing the settlement reiterated that the depository materials must be open to the public. *See* Order Relating to Document Issues Heard on November 17, 1998, State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, slip op. at 2 (Minn. Dist. Ct. Nov. 24, 1998).

3. The Minnesota Select Set, comprising about one to two percent of the total collection, are a subset of documents selected and further processed and analyzed by plaintiffs' attorneys in Minnesota. Whenever a Minnesota plaintiff's attorney designated a document for further study, the defense attorneys were notified of the selection of that document. Other jurisdictions have obtained copies of the Minnesota Select Set on CD-ROM for use in their own cases. In general, plaintiffs' counsel have relied primarily on the Minnesota Select Set, doing little independent research leading to the use of additional documents in their own cases. For example, in the Texas case, the judge ordered that all documents selected in the Minnesota case be turned over to the Texas plaintiffs on CD-ROM, thus reducing their research time considerably. *See* Roberta B. Walburn, *The Role of the Once-Confidential Industry Documents*, 25 WM. MITCHELL L. REV. 431, 433 (1999). The Minnesota Select Set documents are specifically designated at the depository. A description of how to access the Minnesota Select Set at the depository is included in the users' manuals. The same barriers impairing public access to depository information are those which have led to the scarcity of independent plaintiff research. Because of defendants' intransigence in asserting privilege claims, there have not been any significant document accessions through discovery in the non-Minnesota cases.

4. Less than 1000 of those documents were actually introduced as exhibits in the five-month trial. The trial exhibits are available at the Minnesota depository.

On the day the depository opened for public use, one lone researcher, a law student, signed-in to peruse the documents. From that cautious beginning, public usage has increased at a steady rate. Trial lawyers from other states have been the heaviest users, as well as attorneys from other nations. Among the other users were parties to the Minnesota case, reporters, advocacy groups, investigative researchers, academic and private researchers, legislators, and students.

The court ordered the tobacco depository records opened to the public in May 1998.[5] The May 1998 settlement agreement and corresponding consent judgment explicitly require that the defendants keep the Minnesota depository open to the public for ten years.[6] Minnesota's settlement further specified that documents released or disclosed in any other U.S. smoking and health litigation are to be delivered to the Minnesota depository within thirty days of their production.[7] Minnesota has, in effect, become a national repository of tobacco industry documents.[8]

The Minnesota depository, funded by the tobacco industry (and indirectly from document scanning and photocopying fees) is open to the public five days per week. Day-to-day operation of the depository is administered by Smart Legal Assistance, a private legal document-handling contractor. Copying, scanning and shipping is administered by Merrill Corp.

The depository has twelve computer workstations with an electronic search engine that implements public portions of the 4B index.[9] The search engine allows retrieval and designation of

5. *See* Consent Judgment, State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, 1998 WL 394336, at *3 (Minn. Dist. Ct. May 8, 1998).

6. *See id.* The settlement provides for transfer of non-privileged records to the Minnesota Historical Society ("MHS") at the end of the ten-year period. *See id.* Discussions between the MHS and the state of Minnesota regarding the mechanics of such transfer are still in the preliminary stage.

7. *See id.* at *4.

8. In a quite extraordinary development, it appears that millions of additional documents from the files of the now-defunct Tobacco Institute will be maintained by the New York State Archives in Albany, New York. *See* Conversation with Robert Norton, Minnesota Historical Society, in St. Paul, Minn. (Mar. 25, 1999).

9. The index set currently used at the Minnesota depository is called the "4B" index, referencing paragraph 4B in the applicable court order. *See* State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, slip op. at 6 (Minn. Dist Ct. July 14, 1995). The 4B index is the set of defendant indexes made available to plaintiffs and also to the public. It was created by the tobacco industry under the judge's order to provide an index for the millions of documents they were

documents and can segregate tagged documents by individual box number. The documents cannot be retrieved online as images at the workstations; rather the box(es) containing the desired documents must be requested separately.

Both national and locally-based investigative reporters have found the depository material fertile ground for some startling journalism, for example:

- Tobacco companies secretly developed genetically altered strains of high nicotine tobacco in violation of federal law and consent agreements.[10]

- Union leaders across the nation received previously undisclosed lobbying stipends from the tobacco industry.[11]

---

sending to the depository. *See id.* The industry provided a minimal identification of the contents of each of the documents. The 4B index, containing 2.6 GB of information, is available for use onsite and is also available on CD-ROM from Stirewalt & Associates. *See infra* note 46.

The 4B index, while flawed, does contain a vast quantity of useable data. One potential starting point for cataloguing the collection would be to consolidate the various 4B index subsets from each defendant company or entity into a master 4B index. A comprehensive index would reduce the number of searches that a researcher would need to perform by a factor of seven (or by a factor of fourteen if the privilege logs are included).

A consolidated 4B index appears to provide a good cost-effective starting point for greater public accessibility to the depository collection. It is important to note, however, that some researchers will continue to wish to search only one defendant's documents. Therefore, a comprehensive index should still allow searches by individual defendant.

The 4A indexes, created for the use of the defendant counsel, have much more detailed descriptions. The district court ordered that the 4A indexes be made available to the public. *See* Order Relating to Document Issues Heard on November 17, 1998, State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, slip op. at 1-2 (Minn. Dist. Ct. Nov. 24, 1998). The trial court released the 4A index to the public on November 25, 1998. *See id.* The court order was appealed on December 27, 1998, and subsequent appeals are still pending. Those who have seen the index assert that it is more usable than the 4B index. Access to the 4A index would greatly speed the process of evaluating the documents.

Due to the court's recent decision to release the 4A index, and a pending appeal of that decision, this article does not assess the characteristics of the 4A index. It seems plausible to assume that the indexes created for the use of the defendants would be superior to the minimal index created for use by plaintiffs. Public availability of the 4A index would be of great value in any effort to make the depository collection more accessible.

10. *See* Todd Lewan, *Quest for the Nicotine Kick: Inside Big Tobacco's Effort to Develop a More Addictive, Profitable Leaf Called 'Y-1'*, DURHAM HERALD-SUN, Sept. 13, 1998, at F1.

11. *See, e.g.*, Greg Gordon, *Ethics Board to Investigate Teamsters Lobbyist*, STAR

In some cases these payments were sent directly to their homes.[12]

- Tobacco lobbyists arranged contributions to the favorite charities of key legislators.[13] In some cases, the legislators selected the charity and suggested the desired contribution.[14] The industry provided the check to the legislators to give to the charity.[15]

- Tobacco industry internal reports indicate that a Minnesota State Fire Marshal and president of the National Association of State Fire Marshals adjusted in the industry's favor a report on a fatal cigarette-caused fire.[16] Moreover, the industry had a complex system of financial rewards to firefighter organizations across the nation, designed to prevent legislation that would have required cigarettes to meet fire safety standards.[17]

- Federal judges, including former Supreme Court nominee Douglas Ginsburg, accepted luxurious trips to symposiums sponsored by the tobacco industry.[18] Some of these judges later heard tobacco cases.[19]

The depository documents contain evidence of numerous additional matters that are altering the public perception and the legislative and legal status of tobacco in society. However, there are significant barriers to the use of this depository.

TRIB. (Minneapolis-St. Paul), July 14, 1998, at 3B.

12. *See id.*

13. *See, e.g.*, Greg Gordon & Tom Hamburger, *Tobacco Gave to Pet Charities of 3 Legislators*, STAR TRIB. (Minneapolis-St. Paul), June 24, 1998, at 1A.

14. *See id.*

15. *See id.*

16. *See, e.g.*, David Shaffer, *Tobacco, Fire Groups Linked: Companies Opposed Fire-Safe Cigarettes*, ST. PAUL PIONEER PRESS, July 13, 1998, at 1A; *see also* Greg Gordon, *Lobbyists Prove to Be a Powerful Weapon for Tobacco Industry*, STAR TRIB. (Minneapolis-St. Paul), June 29, 1998, at 3B; Myron Levin, *Big Tobacco's Dollars Douse Push for Fire-Safe Cigarettes; Lobbying; Firms Bankroll Experts, Alliances with Safety Groups to Resist Product Changes, Papers Show*, L.A. TIMES, Jan. 1, 1998, at A1.

17. *See, e.g.*, Levin, *supra* note 16, at A1.

18. *See* Tom Hamburger & Greg Gordon, *Tobacco Firm Linked to Travel by Judges, 2 Later Participated in Smoking Cases*, STAR TRIB. (Minneapolis-St. Paul), July 19, 1998, at 1A.

19. *See id.*

III. THE MINNESOTA COLLECTION

As of late 1998, the Minnesota depository held approximately twenty-six million pages containing 3,833,038 documents, stored in about 11,400 boxes.[20] The breakdown by source was:

Philip Morris	1,433,802 documents
R.J. Reynolds	780,825 documents
Brown & Williamson	516,566 documents
American Tobacco	488,110 documents
Lorillard	305,246 documents
Center for Tobacco Research	204,063 documents
Tobacco Institute	104,453 documents

There are also other materials that are not yet indexed within the 4B indexes,[21] such as 170 boxes of Liggett Group documents, about forty boxes from Oklahoma, twenty-two boxes of BAT documents from the Guildford depository (probably part of the Guildford Select Set), and dozens of boxes of material from litigation in other jurisdictions. These documents are not included in the above statistics. There are also several dozen nonindexed boxes of documents that include lobbying documents relating to Minnesota and other jurisdictions. Materials continue to be added to the collection at a steady pace, leading to storage space concerns.[22]

IV. THE EXISTING INDEXING SCHEME'S SHORTCOMINGS

A researcher wishing to do a search on, for example, lung cancer, would need to conduct a minimum of fifteen separate searches within the depository indexes to find out which documents have titles that suggest content relating to lung cancer. First, the researcher would attempt to find keywords relating to the subject, or to individuals who conducted research on lung cancer. Next, the researcher would search each company index. Then she would search the privileged documents[23] logs at the depository, and

20. Conversation with Jay Witthoft, staff member of the Minnesota tobacco depository, in Minneapolis, Minn. (Oct. 1998).

21. *See supra* note 9 and accompanying text.

22. *See generally* David Hanners, *Tobacco Document Storehouse is Already Bursting at the Seams*, ST. PAUL PIONEER PRESS, Feb. 4, 1999, at 1A.

23. The court reviewed, using a modified sampling method, 250,000

finally search the Liggett documents on the Bliley web site.[24] There are also at least 250 boxes of documents for which no searchable index exists. This is not an efficient search system. The successful and conscientious researcher will search interactively because some documents will lead to the identification of additional authors, recipients, keywords or other potential search parameters.

The existing indexing available to the public (the "4B index") is unsatisfactory and unsuitable for the requirements and expectations of most users. While the electronic search software is quite flexible, its capabilities are unrealized due to serious flaws and gaps in the data set. Such flawed data is as expected from the ephemeral and highly adversarial nature of discovery production. While the industry was obligated by court order to provide some sort of index to the truckloads of documents they delivered to the depository,[25] they were under no obligation to provide a useful index with practical and efficient search capabilities.

Each of the 4B corporate indexes contains approximately twenty fields. Among those fields are author, date, document type, copy recipients, document title, Bates stamp number, and recipient.[26] The title field is not often particularly useful. The titles are generally assigned in an arbitrary manner; in many cases the title field is simply left blank. Any of the fields may contain blank, erroneous, or generic filler. Some fields contain entries with erratic or non-standard formats, or input errors. Overly broad title descriptions obscure the true content of most documents. For example, "Shoreview City Council" might be a list of council

---

allegedly privileged documents and released 39,000 of those documents. 211,000 privileged documents remain sealed and are now being withdrawn from the depository. Included in the documents not released were nine categories including youth marketing documents. These were documents that could only be released in different litigation or by the industry. The privileged logs are the indexes for about one million pages. *See* Tom Bliley, U.S. House Commerce Committee Documents (visited Mar. 12, 1999) <http://www.house.gov/commerce/TobaccoDocs/documents.html>. Representative Bliley, who is the chairman of the House Committee on Commerce, subpoenaed the documents and posted them on the House Commerce Committee website ("Bliley website"). *See id.*

24. The Bliley website has about 40,000 documents posted in unsearchable files. *See id.* One private vendor sells a searchable 17 CD-ROM set that contains the documents available at the website.

25. *See* State *ex. rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565, slip. op. at 6-7 (Minn. Dist Ct. July 14, 1995).

26. The indexes also included "attorney comments," which are considered work product and were not sought during discovery.

members, or else a strategy to avoid regulation of underage smoking. In some cases, troubling information appears under overly general notations. For example, startling information about how certain types of ingredients cause cancer may be listed under "research." Such flaws can cause great hardship to researchers accustomed to ordinary archival facilities, whose databases are arranged with the intent of facilitating document location.

In most cases, document content is not obvious from the index description. Documents cannot be readily searched by subject matter. Despite flexible search software, search results are typically of limited use unless the researcher has identified specific people, subjects or organizations of greatest research interest. In most cases, the researcher has no idea what a document includes, and whether it will be of any use, until the researcher retrieves the applicable box(es) and examines the actual document. As a result, research efforts require substantially more time and energy than if a suitable index was available.

V. USE OF THE DEPOSITORY

The limited number of workstations at the depository means that there may not be sufficient space for all researchers.[27] A researcher can use a workstation to identify documents of possible interest from each of the defendant-specific databases. The system prints out a list of selected "hits" from a search within any one of those databases, and sorts the documents by box number. A researcher can request up to three boxes at a time to examine documents. Each document page is stamped with a unique Bates stamp number. Each box contains a printed list of the Bates number ranges within that box. The researcher can obtain documents of interest by writing out the Bates numbers and giving the request to the depository staff, who supply either photocopies or scanned images on a CD-ROM.[28] A researcher can also request documents by Bates number from the index without handling or examining them beforehand, but this proves to be a wasteful retrieval method due to the large number of blank, "junk" or mis-described documents. The depository generally provides

27. To date, this has only occurred a few times.

28. The depository manager has thus been accumulating document scans as researchers request them. The issue of who owns the rights to scanned data has not yet been raised publicly.

photocopies within forty-eight hours or sooner, but may take longer during periods of heavy use.

The depository verifies the integrity of a box (i.e., that it is complete and nothing is missing) after it is returned by a researcher, or even after it is simply sent to the photocopying department. This is a very labor-intensive function. For this reason, boxes are frequently unavailable to a researcher, particularly when the depository is very busy, or when multiple researchers are looking at the same groups of materials. A box may also become unavailable if a discrepancy is noted between the index sheets and the contents of that box. Such boxes are withheld until the depository managers can resolve the inconsistency with the defendant law firms, a process that can take days or weeks. During the first year of public operation, the depository staff has exhibited outstanding cooperation with researchers.

There are many depository procedures and mechanisms that are vestiges of the depository's original (and residual) document management role in litigation.[29] For example, the depository continues to keep a record of which boxes a visitor examines and which documents are copied. Such records are still available to the defendant tobacco companies.

A. *User Observations*[30]

The millions of pages of documents are organized but not easy to access. It is difficult for a researcher to find what she is looking for. Documents are segregated by defendant company source. Although each page is stamped with its own Bates number, some pages contain more than one Bates number and some documents contain multiple Bates numbers from use in prior litigation. Further, the same document may have been submitted by several tobacco companies and/or trade organizations.

Interspersed among the "confidential" documents are many published articles, legislative documents, and copies of documents from external sources that the tobacco industry collected. These extraneous documents, by their sheer volume, conceal significant industry documents under a pile of relatively useless paper.

29. *Cf.* Dick Youngblood, *Firm Nearly Smothered by Tobacco Documents*, STAR TRIB. (Minneapolis-St. Paul), May 10, 1998, at 3D.

30. The following comments are based on the authors' use of the depository and conversations with other experienced users.

Large portions of the collection are inaccessible by searching in the index. First, many of the index entries have fields that contain no data. (Some of the databases are worse than others in this regard, but all contain a substantial fraction of "blank" descriptions.) Some documents lack any description whatsoever. Others lack any index information in significant fields, such as title and date. Second, much of the index contains typographical and other errors created by hasty keypunch input and insufficient data checking. Those errors can and will inadvertently eliminate from consideration a large number of documents if the researcher depends solely on keyword or name searching.

One way an experienced researcher can counter some of the inevitable typographical input errors is to use the "terms index" capability of the database. This is what is sometimes referred to as an "inverted term dictionary": a list of all the words that appear in that particular data field. By clicking on the "terms index" function, a researcher can select not only a particular name or keyword, but also likely misspellings or variants of that name or keyword. Names of individual persons, in particular, are usually spelled or referenced in several ways. While using the terms index helps to mitigate some flaws in the index, one significant limitation to this approach is that it can only search twenty terms simultaneously. Unfortunately, there are sometimes more than twenty variants of a name or other keyword, thus increasing the number of searches that must be performed.

The depository contains numerous duplicate documents because some documents were submitted by different defendants or in different discovery production requests. Unfortunately, those duplicates have different bibliographic descriptions in the index, and it is burdensome to determine whether documents are duplicates solely from the index entries.

Researchers must be creative and resourceful to locate the documents they need. For example, researchers can use varying combinations of document dates, discovery request numbers, document types, and so forth. It is possible, for example, to use the "document type" field to select and then print out a list of all "videotapes" or "invoices" and so on. Or a researcher can print out a list of all "letters" sent during September 1973. The database also allows full Boolean searching.[31]

31. Boolean searching is the use of keywords connected by logical

Another difficulty with the index is that documents typically discuss numerous subject matters, and it is nearly impossible to determine the true content, import, or implication of a document without examining the document. A letter may be described as being from one individual to another, and even if the researcher knows the identity of those persons, the document might concern rerouted political contributions or youth marketing plans, or just a planned holiday retreat.

There are also separate, limited searching tools for the so-called "privileged documents" of each of the seven defendants. As a result, there are fourteen indexes to search for each topic of interest. In practice, this becomes very tedious, and the researcher must be extremely careful to record the precise searches that she has performed.

B. User Expectations

The tobacco depository differs from most research archive collections. Research normally involves the task of reviewing, filtering and scrutinizing large masses of data to locate useful information. Normally, an archive puts the burden of filtering that material on the user. Typically, an archival researcher expects to do that legwork. The archive is expected to provide the finding aids, tools and rudimentary direction to enable the researcher to learn how to find helpful information.

Those who use the tobacco depository seem to have different expectations. Given the sheer size of the collection and its lack of hierarchical structure, it is unrealistic to expect even the most tenacious researchers to go through these records on their own. It could take months or even years for an individual researcher to find the desired records. As a result, researchers become frustrated when their basic expectations are not met, and the material sought is not readily accessible.

C. Policy Dictates

On July 17, 1998, President Clinton issued a memorandum to the Secretary of Health and Human Services ("HHS") that highlights the importance of increasing the accessibility of tobacco

expressions such as AND, NOT, OR and other various arithmetic operators to construct a desired search strategy and locate specific items.

industry documents available as a result of litigation and congressional subpoena.[32] Citing the potential value of these documents to the public health community, the memorandum directed the Secretary to:

1. Propose a method for coordinating review of the documents and making available an easily searchable index and/or digest of the reviewed documents;
2. Propose a plan to disseminate widely the index and/or digest as well as the documents themselves, including expanded use of the Internet; and
3. Provide a strategy for coordinating a broad public and private review and analysis of the documents to gain critical public health information. Issues to be considered as part of this analysis include: nicotine addiction and pharmacology; biomedical research, including ingredient safety; product design; and youth marketing strategies.[33]

The memorandum provided for a ninety-day deadline (by mid-October, 1998) for the HHS secretary to submit a plan designed to accomplish these goals.[34] The HHS secretary submitted the plan, the details of which are not publicly available at the present time.

VI. CATALOGUING STRATEGIES: OCR INDEXING VERSUS SUBJECT INDEXING

There is a tug of war between two primary cataloguing strategies. One approach would establish an expert consensus of what constitutes a useful document, extract the most useful documents from the mass of "junk" documents, and catalogue all "useful" documents using appropriate subject terms. The second approach envisions imaging (scanning) all (or most) documents,

32. William Jefferson Clinton, *Public Availability of Tobacco Documents*, Executive Memorandum from the President to the U.S. Secretary of Health and Human Services (July 17, 1998).

33. *Id.*

34. *See id.*

with minimal pre-screening, applying optical character recognition ("OCR"), compiling a full-text searchable database, thus permitting dumping of image and text data onto a researcher's hard drive to analyze anywhere in the world. The remainder of this paper describes and evaluates both of these approaches, and the procedural issues expected from implementing either approach. This paper also suggests how both approaches could be combined into a hybrid cataloguing strategy using the advantages of both types of cataloguing.

A. *Sorting into Record Groups*

Optimally, a depository collection of this magnitude would be maintained in distinct record groups and subgroups, with related materials stored in proximity.[35] The huge size of the collection, and the quasi-random ordering of the material, makes physical reorganization impracticable. Such a project would also remove the depository from public access for an unreasonable period of time.

B. *Cataloguing All Documents Versus Selected Documents*

The first issue is whether to "catalogue" all of the documents or only a selected set of documents. Cataloguing all of the documents might be advantageous in understanding certain diffuse industry relationships and would create a truly comprehensive database for researchers. In addition, comprehensive cataloguing, by definition, would reduce the need for highly skilled personnel to cull out insignificant documents.

At first glance comprehensive cataloguing appears impractical within reasonable time and budget constraints. However, the use of high capacity scanners, OCR software, and modern data storage make comprehensive cataloguing feasible. Modern scanners can scan sixty pages per minute, or over 3500 pages per hour, at a cost of ten to twenty-five cents.[36] Running twelve scanners at once, the

35. Most archival collections are assembled and guided by records retention schedules, established before the creation of the documents themselves. In this case, the materials already exist, and the collection grew through accretion over time.

36. Generally accepted costs for scanning using bulk sheet feeders. Factors that can reduce this scanning rate and increase the cost are curled paper, stapled reports and irregular sheet sizes. Contractor quotes are likely to be higher than in-house scanning.

twenty-six million Minneapolis depository pages could be scanned in a month.

Once preliminary setup has been accomplished, the OCR process can be automated to an extremely high degree. OCR and scanning can be performed in parallel: as soon as a document is scanned the image can be routed to OCR. OCR utilizes substantial computer processing time, but the highly automated nature of the OCR process requires relatively low employee staffing levels. OCR would be able to recognize text from an estimated eighty percent of the depository documents. The remaining documents would likely require additional human inspection. Following OCR, the resulting text would be compiled into a database to allow the creation of a searchable index, again a fairly automated process.

The other option is presorting the documents to require cataloguing of only a desirable (useful) portion of the collection. The advantages of partial cataloguing include filtering out much of the extraneous material of little or no interest to researchers at the beginning of the process. This extraneous matter includes "junk matter" such as photocopies of blank sheets, multiple duplicative pages, repetitive matter, and so on. There are also large quantities of otherwise publicly available material such as newspaper clippings, magazine articles, mark-ups of legislative bills, and so on. The authors estimate that nearly one-fourth of the depository materials fall into this category. Weeding out one-forth of the documents would have the additional benefit of reducing logistical costs for subsequent processing steps, and benefits future users by increasing the fraction of useful material.[37]

Standard archival practice does not typically include cataloguing every item in a collection.[38] Instead, sound archive management establishes a hierarchy of record series, subseries, groups, subgroups, and so forth.[39]

There are both short-term and long-term needs. In the short-term, immediate public health concerns would benefit from rapid cataloguing relying primarily on technology and requiring less skilled human review. In the long-term, a subject-based added-

37. However, increased document handling, such as the need to refile the culled documents back into the boxes, may reduce this savings or even result in increased costs.

38. *See* Telephone interview with a staff member at the Minnesota Historical Society, in St. Paul, Minn. (Sept. 18, 1998).

39. *See id.*

value catalogue would allow systematic study in a number of sociological, historical and public policy areas. Thus, both approaches have a place in depository planning.

The authors believe that a well-briefed and conscientious archivist can readily discern material suitable for indexing from material that is not worth indexing.[40] Such separation of documents is a simple process—any documents of doubtful utility would be retained and scanned. Only documents that are obviously worthless according to simple predetermined criteria would be bypassed. The authors believe that such criteria would not be difficult to develop.

C. *Full Text Searchability Versus Subject Term Searching*

The extreme limit of indexing would be to process all twenty-six million pages in the depository into a full-text searchable database. At first glance that appears to be an intimidating task that would require enormous resources. However, the continuing exponential improvements in computer processing hardware and document handling software have made such an approach eminently feasible.

Dozens of commercial search engines could be applied to locate information within the completed text database.[41] Faster and better text search engines are becoming available all the time, and the text database would always allow searching by different search engines depending upon the research need.

According to one approach, searching by assigned subject terms makes a lot more sense for such a large and varied database. Subject matter cataloguing is generally more efficient for most types of searches than simple keyword cataloguing. For example, the West Key Number cataloguing system for legal opinions is usually more efficient than a pure keyword search at directing the legal researcher toward the most relevant material with the least

40. If a highly skilled worker can review 20 pages per minute for first-order culling decisions, three or four reviewers would be required to keep up with each document scanner.

41. It is envisioned that the finished text database would be stored in a .TXT or ASCII format for generic searching capability not wedded to proprietary platforms. *But see* Communication with Michael Tacelosky, (Feb. 28, 1999) (suggesting instead that storage in XML or some other agreed-upon document mark-up language). Mr. Tacelosky, a private individual, has scanned, subjected to OCR, and indexed hundreds of thousands of documents. *See id.* (referencing <http://www. tobaccodocuments.com>).

extraneous matter in most cases.[42] One reason for this is that many keyword terms have synonyms, variants, or euphemisms[43] that a researcher is less likely to use, but that may be the sole connection to a particular document of interest. Another reason is that common keywords can retrieve huge numbers of records of questionable relevance to the desired subject matter.

On the other hand, one difficulty with subject matter cataloguing is the need for highly skilled personnel to examine every document in order to code it with the proper designation. Moreover, it is necessary for the indexers to use a common set of subject terms, groups, and keywords, and thus to have some consensus on which subjects and descriptors should be chosen for use in cataloguing the collection. Many of the documents would require many descriptors to do the document justice, as many of the documents address various topics simultaneously. Further, the time to catalogue the collection adequately by subject matter could unduly delay public access to the collection.[44]

The 4B indexes include a "request number" field, enumerating the discovery requests served on the defendants.[45] A careful researcher can use that field for limited subject matter searching for discovery requests involving a narrow subject matter description. For example, one of the request numbers identifies most Minnesota lobbying documents. However, the request number field is of limited value for two reasons: in many instances discovery requests were broad, and the companies themselves coded the documents to discovery request numbers. Further, the description of the discovery requests is unclear in the manuals that presently accompany the indexes.[46]

---

42. The West Key Number system was first implemented with the publication of the West Publishing Company's First Decennial Digest (1897-1906), as a way of organizing and accessing large masses of legal information before the advent of computers and other data-processing equipment.

43. As an example, some tobacco industry scientific papers used arbitrary code words to refer to troubling concepts. For example, the word "zephyr" represented the word "cancer" in some BAT documents.

44. At a rate of one minute to code each document, an indexer could code 480 documents in an eight-hour day. At this rate, a team of 20 indexers could code the roughly three million depository documents in 312 working days.

45. Documents discovered from plaintiffs are not available at the tobacco document depository.

46. *See* ROBINS, KAPLAN, MILLER & CIRESI, TOBACCO: A CATALOG OF MINNESOTA CASE RELATED MATERIALS. Plaintiffs' counsel, Robins, Kaplan, Miller & Ciresi, L.L.P., has published a catalogue of materials compiled as part of *State ex rel.*

There are several tools that may prove useful in indexing the collection. Several organizations have well-developed thesauri that could be modified for use with this project, including the Office on Smoking and Health ("OSH") at the Centers for Disease Control and Prevention and the National Library of Medicine at the National Institutes for Health. The National Archives and Records Administration and the Library of Congress also have useful bibliographic control resources. The OSH Thesaurus provides a good base for depository indexing, but the indexing authority must supplement this technically oriented thesaurus with additional terms to address the political/lobbying overtones of many of the documents.

D. Scanning and Optical Character Recognition

While image scanning is well accepted as a way to gather and sort large quantities of information, no single image file format is universally accepted. The authors believe that the key factors in selecting an image format are: ease of researcher retrieval, speed of researcher retrieval, format transparency, and compatibility with common software.

Great strides have been made in recent years in the area of optical character recognition. OCR involves scanning in a page of text, or mixed text and graphic materials, automatically extracting the printed textual portions, automatically identifying the numbers and letters, translating those characters into an ASCII data set, and creating and storing the document so that the textual information can be recreated or analyzed.

It has been suggested that the OCR process applied to depository documents will result in a high number of errors.[47]

Humphrey v. Philip Morris. The catalogue includes indexes to pleadings and court orders, a list of defendants' and defendant employees' depositions, expert witness depositions, and jury instructions. *See id.* The catalogue also contains a list of admitted trial exhibits offered by both parties, a description of document foundation materials, parties' experts' reports, 4B indexes, defendants' privilege logs, and databases identifying individuals appearing on the defendants' privilege logs. *See id.* All of the items are available on CD-ROM from Stirewalt & Associates, 25 East Serpent Road, P.O. Box 416, Deerwood, MN 56444, telephone: 800-553-1953. In addition, the 4A indexes may also become available to the public pending defendants' appeal of the court's order of November 25, 1998 to release those indexes. *See* State *ex rel.* Humphrey v. Philip Morris Inc., No. C1-94-8565 (Minn. Dist. Ct. Nov. 25, 1998).

47. Computer printer output, multi-generation photocopied documents,

Without special precautions, OCR of documents of varying quality can result in unacceptable error levels and impede reading of the resultant textual output. However, error levels that might be a problem in short documents are of less significance in a huge database of this type. First, it is unlikely that errors will appear in *significant* keywords. Second, even in the event significant words are lost, those keywords will likely appear elsewhere in the document, or will be found by alternative search strategies.[48] Thus, a massive database with a moderate error rate but a high degree of internal redundancy can still be extremely useful and productive. Once this database (text and images) is disseminated widely, any future analytical techniques can be applied to study the data as those techniques become available.

In some cases, it is not the typewritten or printed text that is most significant. Rather, the significance lies in the handwritten notations ("marginalia") and the origin, authorship or location of those notations. With preplanning, the OCR process can identify and tag documents containing handwriting marginalia either for closer human examination or scanning at a higher resolution. Some material is graphical in nature and would not translate via the OCR process. Graphical documents can also be separately identified and tagged, while many documents containing both mixed text and graphical image are identifiable and accessible through the text indexing process.

The authors estimate that the full document text would occupy approximately 52 gigabytes "GB," and that the complete set of scanned document images would require about 1.69 terabytes, or 1690 GB.

The authors believe that scanning and OCR have proven their worth and may well be suited to making depository information resources available with minimal delay.

"noisy" documents and mixed image and text are all handled with varying degrees of success by contemporary OCR software.

48. The OCR software can be set up with a glossary of the most important keywords and terms to allow the program to place higher priority on recognition of those words in the event of recognition errors.

E. *Image Resolution*

The authors believe that imaging of the documents should use a resolution of between 300 and 400 dots per inch (dpi). Lower resolution may make it more difficult to see details, particularly on documents of lower quality or small print. Higher resolution may be slightly clearer for printed documents but add little to reading on a computer monitor, require significantly more memory, and is not necessary for OCR purposes.

F. *Internet Access*

Until now, the Internet has been an unsuitable medium through which to utilize the tobacco depository documents. The main problem is that it takes an unacceptable length of time to download and view documents retrieved via the Internet. Even brief download delays are unacceptable because a researcher must be able to examine documents quickly to avoid frustration, achieve minimal productivity, and achieve anything useful.

Rapid document examination is required because this document collection is not comprised of *individual* or *self-standing* documents or visual works of public significance, like the Declaration of Independence, the U.S. Constitution, the Magna Carta, the Mona Lisa, etc. Rather, the collection has value based on the *interrelationships* among documents and the capacity to follow lines of communication and activity through the tobacco industry and beyond. This kind of comparison and study requires the examination of thousands of pages within a short period of time—until recently, an impractical approach via the Internet. Despite these impediments, many tobacco industry documents have been placed on the Internet.[49]

---

49. In September of 1998, the HHS released the following summary of tobacco documents on the Internet:

1. U.S. House Committee Documents:
   Released December 18, 1997
   http://www.house.gov/commerce/TobaccoDocs/documents.html
   Website, CD-ROM available through GPO, commercial vendor
   First Release December 18, 1997, approximately 800 documents
   Second Release April 22, 1998, approximately 39,000 documents
   Third Release June 18, 1998, approximately 400 documents
2. Brown and Williamson Collection:

The recent availability of high-speed data access technology (i.e., ADSL, ADSL-Lite, HDSL, fiber optic cable, and coaxial cable) makes Internet access to scanned depository documents quite practical. Newly available and affordable high speed telephone lines for home access would require about one half second to download a document page, a tolerable time delay that a user will barely notice. High-speed institutional access lines would result in nearly instantaneous access.

G. CD-ROM Distribution

CD-ROMs (and their successor optical disk products, Digital Versatile Disks ("DVDs")) have the potential to distribute information quickly and in a cost-effective manner to researchers interested in specific portions of the archive. CD-ROMs are inexpensive, have well-established standards for reading, and are reliable as a storage medium.[50] Each CD-ROM can store up to 200,000 pages of text, or about 5800 images. DVDs hold several times that capacity. Newer versions of optical disk technology being put on the market this year can store even more information. A large shoebox of DVDs would hold the entire tobacco depository image collection; a mere handful of DVDs (or a shoebox of CD-ROMs) would hold the entire text database.

While CD-ROMs may be a good way to duplicate and transport large quantities of data, they are probably not the best way to *use*

---

http://galen.library.ucsf.edu/tobacco/bw.html
Website, CD-ROM
approximately 2500 documents

3. Tobacco Resolution Site:
http://www.tobaccoresolution.com/
number of documents: unknown
also: http://www.philipmorris.com
also: http://tobaccoinstitute.com
also: http://documents.rjrt.com
also: http://www.lorillarddocs.com
also: http://www.bw.aa.latg.com
4. Blue Cross, Tobacco Litigation:
http://www.mnbluecrosstobacco.com/toblit/trialnews
39,000 documents

Additionally, a private individual has scanned, subjected to OCR, and indexed hundreds of thousands of documents. *See* Communication with Michael Tacelosky, (Feb. 28, 1999) (referencing <http://www.tobaccodocuments. com>).

50. *See, e.g.*, Arthur Tolsma, *Recordable CD's Reduce Paper, Hard-Disk Storage; CD-ROM Is Rapidly Becoming a Preferred Medium for Record Archives and Web Data Storage*, NAT'L L.J., Feb. 3, 1997, at C9.

the data. Rather, the files should be dumped onto large hard drives for rapid access. Sixteen GB hard drives are becoming very affordable and simple.

A paradigm shift is occurring. No longer will the researchers be required to travel to the depository. By either CD-ROM or by high-speed Internet access, the data is brought to the researchers rather than the researchers having to go the data. Nevertheless, this scenario still requires significant up-front planning, data processing, and weeding. This process would also benefit greatly from enhancement such as subject matter coding.

H. Subject Matter Classification

Subject matter classification would add substantial value of the depository collection. The physical reorganization and sorting of documents by subject matter is impractical due to the size of the collection. However, such sorting could be accomplished in the virtual world, by "moving" electronic representations of documents into useful virtual collections.

Documents within a subject matter class can be catalogued further within that class. An entire class of documents might be published on one or more DVDs or CD-ROMs.[51] This approach permits great flexibility in the allocation of cataloguing resources. For example, if marketing to youth was a subject area of great interest to the research community, that material could be indexed in more depth. Whereas, if product sampling was of lesser interest, that material could be scanned and the images placed onto CD-ROM for further study as required. Subject classes of lesser interest could be identified and basic indexes published for reference purposes.[52]

The HHS has cited twelve general categories as being of interest from a public health perspective. Among those are

51. The tobacco depository has placed as much as 4000 pages of scanned material onto a single CD-ROM. It is estimated that 40,000 pages or more of scanned images could be placed on a DVD.

52. The National Security Archive ("NSA"), a private archive located in the Washington, D.C., area takes a productive approach toward making sense of huge masses of unsorted records. The NSA was established by investigative journalists to preserve history by making formerly classified government documents available to scholars and researchers. Typically, the NSA staff collects and assesses an entire class of documents, and then publishes an analytical text for public sale on that topic. In this way, both the raw documents as well as the finished analysis are accessible to the research community.

science, attorney-related involvement in smoking and health, public statements made by a defendant or by the industry regarding smoking and health, documents referring to children, documents relating to government regulatory activity, and documents relating to patents or the EPA.[53] These categories are quite broad and do not necessarily exhibit the level of specificity that the ultimate categorization scheme would encompass.

Subject matter categories would allow researchers to study documents relating to such interdisciplinary topics as: nicotine pharmacology, nicotine addiction, health consequences of tobacco use, tobacco product additives, tobacco product design and manufacturing, product packaging, advertising and promotion, agriculture, marketing research, disruption of public health programs and intervention activities, manipulation of scientific processes, environmental tobacco smoke, and policy research.[54]

The cataloging agency could create a draft subject matter classification scheme. This is not an insurmountable task. It would be beneficial to receive input and comments from the user community, and in this way the research community could reach some consensus on the subject matter indexing system to be applied.

I. Consensus on Subject Matter Classification

Due to the interdisciplinary nature of these documents and the large number of constituencies represented by depository researchers, it would be advisable to establish an advisory group to guide subject matter classification. This group would include attorneys, scientific and medical researchers (including those specializing in addiction, biochemistry/genetics and agriculture), tobacco control specialists, public health officials, and representatives from the academic community, political activists, and news media. It is *most* critical that the group includes archivists, historians and experts in library science to lend the benefit of archival experience. The advisory group would convene for a series of four daylong conferences to establish the minimum and ideal requirements for a system to meet their needs. The

53. *See* U.S. Department of Health & Human Services, Address at Technical Assistance Meeting on Retrieving, Cataloguing and Analyzing Tobacco Industry Documents (July 1998).

54. *See id.*

group would solicit nationwide input on which areas are of interest to researchers, and use this information as a guide to establishing a classification system. Although this is a novel approach, it would appear to meet the new problems and opportunities presented by the depository materials.

VII. CONCLUSION

A large number of tobacco depository documents have moved into the public domain as a result of court orders and discovery activity during recent litigation. Study of individual documents has led to media exploration of the political, social and medical effects of tobacco. Effective societal use of these records and the information contained therein hinges on systematic indexing methods, taking into consideration the needs of anticipated end-users.

Releasing and cataloguing tobacco industry documents will only be useful if they undergo systematic analysis, resulting in publication in the scientific and lay literatures.[55] The mountain of documents in the depository contain, among other things, tobacco industry research regarding additives, addiction, health effects, political strategies, internal debates and lists of secret allies. The inner workings of what may prove to be the most powerful and corrupt industry in the nation are documented as needles within a haystack of paper. Governments, researchers, legislators, public health activists, attorneys, reporters and the general public ought to know what the documents say. Under the current indexing system, only the most tenacious will have access.

To create a depository indexing system consistent with user expectations, users must be allowed to quickly locate all documents with respect to a desired subject, while excluding the maximum number of unrelated or irrelevant documents. The existing depository indexing system is unsuitable for that purpose.

It may be advisable to weed out documents of negligible research value, given the large fraction of such material. One indexing option, optical character recognition, can turn the depository's unwieldy mass of uncorrelated materials into a searchable set of full text files. Scanning and text recognition of this type is now achievable given recent advances in computer

55. ROSWELL PARK CANCER INSTITUTE, PROCESSING AND MANAGING TOBACCO INDUSTRY DOCUMENTS 2 (Nadine-Rae Leavell, project coordinator, 1998).

technology.

Another option, subject cataloguing, would allow selection and classification of documents into manageable groups. Such processing would be resource-intensive but add significant value to the collection.

In either case, contemporary optical disk technology can store and transmit both document text files and document images in a surprisingly compact manner. Once transferred to local hard drives, access time for images and text is extremely rapid. High-speed phone line technology will allow similarly rapid access time from remote locations.