Technical Recommendations for Preserving Industry Documents Disclosed in Litigation

UCSF Industry Documents Library
July 26, 2021

Summary

Attorneys General and private parties are engaging in ongoing litigation about tobacco (including e-cigarettes), opioids, global warming and other issues in which important documents and other evidence is being produced. Settlements to date in e-cigarette (Juul) and opioid cases have included key provisions to provide for public access to the discovery materials. To make these provisions a reality, it is important that these materials be provided in forms that can be efficiently made freely available to the public and maintained at minimum cost over the long term.

"Discovery materials" can take many forms, including paper and digital documents, oversized records (such as posters and visual displays), multimedia records (such as audio and video recordings), and three-dimensional objects (such as sample products). The key to providing widespread economical public access is storing the discovery materials in digital form to the greatest extent possible. Any agreement to make discovery materials available should not only deal with such digital or digitizable documents, but also all other types of discovery materials produced.

The UCSF Industry Documents Library, based on two decades of experience collecting, preserving, and providing public access to industry documents disclosed in litigation, offers the following recommendations on how to make these materials freely available in perpetuity and what costs should be included as part of settlements or judgements.

1. Documents produced from an eDiscovery platform should be exported in three formats: native files, TIFF images, and PDF files. If paper documents or other physical materials are produced they should be organized by Bates number or other control number, and sufficient funding should be provided to cover costs of digitization and/or storage.
2. Detailed metadata should be provided for each document, as specified below; if metadata that meet the required standards are not provided by a company, costs should be included in the settlement to cover the costs of creating high quality metadata.

3. Optical Character Recognition (OCR) should be performed on digital and digitized documents to generate raw text with page-break indicators, which can be used for full-text search, or for screening and redacting any protected information (if required).
4. Specific limited provisions should govern document redaction, including the creation of a redaction log which indicates the type of information which has been redacted, with sufficient detail to allow an assessment of the merits of the privileged, trade secret, or privacy assertion by an independent agent with the authority to resolve any disputes.
5. A procedure should be established by which members of the public may challenge the appropriateness of a redaction or withheld document(s) and appeal to have that document(s) reviewed and released by an independent agent.
6. If a company does not provide metadata, PDF files, OCR text, or perform specified redactions in a timely manner, additional funding must be provided to enable a documents repository to do this work.
7. Funding to process documents and maintain long-term free public access should be included in the settlement or judgement.

Background

The UCSF Industry Documents Library (IDL) is a digital archive which provides public access to more than 15 million documents (94 million pages) from tobacco, opioid, pharmaceutical, chemical, food, and fossil fuel industries released through litigation and other sources.

IDL was established as the Legacy Tobacco Documents Library in 2002 at the University of California, San Francisco (UCSF) for the purpose of preserving and providing public access to 40 million pages of tobacco industry documents released by the 1998 Master Settlement Agreement between the major tobacco companies and 46 U.S. states, 5 U.S. territories, and the District of Columbia. The Legacy Tobacco Documents Library was created with $15 million from the American Legacy Foundation (now Truth Initiative) which also supported the creation of the UCSF Center for Tobacco Control Research and Education (CTCRE). Of this amount, $2.5 million was a 5-year grant, $2.5 million was to cover capital costs of creating the Library and Tobacco Center and $10 million was to create an endowment to cover ongoing costs. Half of these funds were allocated to the costs of creating and maintaining the Tobacco Documents Library.

In 2011 the US Department of Justice negotiated a consent order with the defendants in *U.S. v. Philip Morris* in which the tobacco companies provided an additional $6.9 million to UCSF to cover additional costs of processing and housing tobacco industry documents

disclosed after the Master Settlement Agreement.[1,2] In *U.S. v. Philip Morris* the Department of Justice sued several major tobacco companies for fraudulent and unlawful conduct under the Racketeer Influenced and Corrupt Organizations Act (RICO). The 2006 court order required that the tobacco companies make public all documents produced in litigation related to smoking and health until September 2021, and UCSF has continued to collect all documents produced under the RICO judgment.[3] Additional funds have come from foundation and government grants, but the funding from the MSA (indirectly) and RICO judgment provide the core funding for the collection.

The Legacy Tobacco Documents Library became the Truth Tobacco Industry Documents in 2015 (reflecting the American Legacy Foundation's name change to Truth Initiative) and is now managed under the umbrella of the UCSF Industry Documents Library.

In addition, UCSF also collects documents created by other industries which impact public health – specifically drug (including opioids), chemical, food, and fossil fuel industries. These collections have been funded by a variety of sources.

Recommendations

1. **Documents produced from an eDiscovery platform should be exported in three formats: native files, TIFF images, and PDF files. If paper documents are produced, they should be organized by Bates number or other control number, and sufficient funding should be provided to cover costs of digitization.**

UCSF can accept and process paper documents but doing so adds substantially to processing costs. The fact that most if not all documents now produced in litigation are already digital means that obtaining digital copies will substantially speed processing and lower costs. These digital records are, however, in a wide variety of file formats. These records include word processing documents, PDFs, email messages, spreadsheets, slide presentations, websites, images, audio and video recordings, social media, data files from chat communication platforms such as Slack, and other ever-evolving formats.

Each of these file formats have specific digital preservation issues which must be considered (for just one example, how to preserve tracked changes in a Word document). Digital archivists and other experts in the U.S. and around the world have conducted

[1] (Order #27 Remand: Consent Order Between the United States, the Public Health Intervenors, Philip Morris USA Inc., Altria Group, Inc., and R.J. Reynolds Tobacco Company Concerning Document Disclosure Obligations Under Order #1015, 2011)
[2] (Fernandez, 2011)
[3](Public Health Law Center, n.d.)

extensive research and provided detailed preservation recommendations for many of these formats.[4]

Fortunately, today these discovery materials are usually handled through an eDiscovery software system, which provides the option for files to be exported in various formats: native files; single-page TIFF (Tagged Image File Format) images; or PDF files.

The recommendations below assume that documents will be produced from eDiscovery software. It this is not the case, please contact us and we will provide more specific recommendations based on the formats of the available documents.

We recommend that digital documents be produced in all three formats: native format, TIFF image, and PDF.

Each of these formats has specific advantages and disadvantages:

Type	**Advantages**	**Disadvantages**
Native format	- Original content with significant properties maintained (e.g., track changes, spreadsheet formulas, email headers)	- Dependence on specific software - Can be altered by a user - Some formats not easily viewable in a web browser
TIFF image	- All documents are in a standardized format - Stable, well-documented, widely adopted, and uncompressed file format used for preservation - Supports Optical Character Recognition (OCR)	- Potential loss of original/significant properties - Produced as single pages which must be recombined to form complete document - Larger file size
PDF file	- Stable, flexible format which can be easily viewed, printed, or downloaded	- Potential loss of original/significant properties - Difficult to accurately convert some native formats to PDF (e.g., spreadsheets)

[4] Digital preservation standards have been developed by the U.S. National Archives and Records Administration (The U.S. National Archives and Records Administration, 2019); the Library of Congress (Library of Congress, n.d.); the Digital Preservation Coalition (Digital Preservation Coalition, n.d.) the University of California Libraries (Schaefer, et al., 2020) and many other organizations.

Together, these three formats provide a full package of data which supports preservation of original content, creation of OCR text for document screening/redaction and full-text search, and flexible online access and delivery.

2. **Detailed metadata should be provided for each document, as specified below; if metadata that meet the required standards are not provided by a company, costs should be included in the settlement to cover the costs of creating high quality metadata.**

Each document should be described with the metadata fields listed in the Metadata Specification below so that it is discoverable among millions of other documents. We have found that the quality and quantity of metadata provided by a company can vary widely, with some documents missing such basic information as title, date, or author. To minimize costs, it is important that the settlement specify the specific metadata to be produced for each document.

Alternatively, missing metadata can be created by trained indexers supported by automated tools where possible, but the additional cost (detailed below) can be substantial. Settlements should carefully address this issue and, if necessary, include specific funds for the UCSF Library (or other archive) to create the metadata needed to make the collection useful to the public.

METADATA SPECIFICATION FOR E-DISCOVERY DOCUMENTS

FIELD NAME	FIELD DESCRIPTION
BEGDOC	Beginning Bates number (production number)
ENDDOC	End Bates number (production number)
BEGATTACH	First Bates number of family range (*i.e.,* Bates number of the first page)
ENDATTACH	Last Bates number of family range (*i.e.,* Bates number of the last page of the last attachment)
ATTCOUNT	Number of attachments to an email
ATTACH	Populate parent records with original filenames of all attached records, separated by semi-colons
CUSTODIAN	Name of person from whose files the document is produced
AUTHOR	Author of the e-doc or attachment
RECIPIENTS	Recipients of e-doc
FROM	Sender of email
TO	Recipient of email
CC	Additional recipients of email
BCC	Blind additional recipients of email
FILESIZE	Size of the file
PGCOUNT	Number of pages in the e-doc

FIELD NAME	FIELD DESCRIPTION
DATERECD	YYYYMMDD Date email was received
TIMERECD	[hh]:[mm]:[ss] Time email was received
DATESENT	YYYYMMDD Date sent
TIMESENT	[hh]:[mm]:[ss] Time sent
CRTDATE	YYYYMMDD Date created
CRTTIME	[hh]:[mm]:[ss] Time created
LASTMODDATE	YYYYMMDD Date last modified
LASTMODTIME	[hh]:[mm]:[ss] Time last modified
TITLE	Title field value extracted from the properties of the native file
MODBY	Name of person(s) who modified e-doc
SUBJECT	The value in the subject field of an e-doc or e-attachment
FILENAME	The full name of the native file
DOCUMENTTYPE	The category of document (*e.g.,* letter, email, memo, report, presentation, advertisement, etc)
NAMED INDIVIDUALS	Individuals named in the document who were not authors or recipients
NAMED ORGANIZATIONS	Organizations named in the document who were not authors or recipients
BRAND	The name of any brand or products discussed in the document, if any (*e.g.,* JUULpod, JUUL Device)
PROJECT NAME	Name of any project associated with the document
FILE EXT	The extension of the file
MD5HASH	MD5 Hash Value created during processing
FULLPATH	File source path for all electronically collected documents, which includes location, folder name, file name, and file source extension
RECORDTYPE	Should contain the value of email, e-doc, or e-attachment
APPLICATION	Name and version of the application used to open the file
VOLUME	Production volume number (*e.g.,* V001, V002, etc)
COMMENT	Values extracted from comments metadata field
ENTRYID	Unique identifier of emails in mail stores
ATTLIST	List of each attribute on a previous defined element definition with an DTD
FAMILYDATE	YYYYMMDD Date value of parent file (email or e-doc)
REQUESTNO	Reference number of the specific discovery request for which the document was produced
NATIVELINK	The full path to the produced native on the production deliverable
TEXTPATH	The full path to the produced text files on the production deliverable
CASE	Eight-digit ID number and/or name of the court case for which a document was produced
COURT	The name of the court where the document was filed

FIELD NAME	FIELD DESCRIPTION
EXHIBITNUMBER	Identifier for documents listed as trial exhibits
DATEPRODUCED	YYYYMMDD Date on which document was produced or transcript was received in litigation
COUNTRY	The primary country or countries mentioned in a document
LANGUAGE	Language a non-English document is written in
RESTRICTIONS	Privilege, trade secret, contains redacted material, or none

METADATA SPECIFICATION FOR PAPER DOCUMENTS (AND OTHER DISCOVERY MATERIALS)

FIELD NAME	FIELD DESCRIPTION
DOCUMENTID	Bates Number or other identifying number or alpha-numeric code assigned to a document
MASTERID	A range of Bates Numbers identifying a group of documents found attached to, or physically close to, each other during the discovery process
OTHERNUMBER	An identifying number or alpha-numeric code assigned to a document, in addition to its Bates Number
TITLE	The title of the document
DOCUMENTDATE	YYYYMMDD The date, if any, which appears on the document
DOCUMENTTYPE	The category of document (*e.g.,* letter, email, memo, report, presentation, advertisement, etc)
PERSONATTENDING	Any person present at a meeting mentioned in a document
PERSONAUTHOR	The author of the document
PERSONRECIPIENT	The recipient of the document
PERSONCOPIED	The person(s) copied on a document
PERSONMENTIONED	The person(s) mentioned in the document
ORGANIZATIONAUTHOR	The organizational author of the document
ORGANIZATIONRECIPIENT	The organization(s) which received the document
ORGANIZATIONCOPIED	The organization(s) copied on a document
ORGANIZATIONMENTIONED	The organization(s) mentioned in the document
ORGANIZATIONATTENDING	Any organization present at a meeting mentioned in a document
PHYSICALATTACHMENTS	Document IDs of any documents which are physically attached
FILENAME	If document has been digitized, filename of the scanned digital copy

FIELD NAME	FIELD DESCRIPTION
BRAND	The name of the brand(s) or product(s) mentioned in the document
PAGECOUNT	Number of pages in the document
CASE	Eight-digit ID number and/or name of the court case for which a document was produced
COURT	The name of the court where the document was filed
EXHIBITNUMBER	Identifier for documents listed as trial exhibits
DATEPRODUCED	YYYYMMDD Date on which document was produced or transcript was received in litigation
AREA	The physical location where a document was found in the offices of the providing company
BOX	Box number where the physical document is stored
FILE	The title of the file folder in which a document was originally kept
COUNTRY	The primary country or countries mentioned in a document
LANGUAGE	Language a non-English document is written in
RESTRICTIONS	Privilege, trade secret, contains redacted material, or none

3. **Optical Character Recognition (OCR) should be performed on digital and digitized documents to generate raw text with page-break indicators, which can be used for full-text search, or for screening and redacting any protected information (if required).**

It is very important that digital documents, whether provided in native, TIFF, or PDF format, are accompanied by text (TXT) files containing the raw text of the file. Raw text is required to conduct text analysis to identify and locate protected information which must be redacted; it is also necessary for providing full-text search and for text mining or other computational research.

Text files should include page break indicators so that specific text can be located on a particular page of the corresponding native file. If text files are not provided, they can be created from TIFF images or PDF files, at an additional cost (detailed below) which should be included in the settlement payments.

4. **Specific limited provisions should govern document redaction, including the creation of a redaction log which indicates the type of information which has been redacted, with sufficient detail to allow an assessment of the merits of the privileged, trade secret, or privacy assertion by an independent agent with the authority to resolve any disputes.**

These redactions should be limited to:

- Confidential Personal Information and personnel files, including home addresses, phone numbers, Social Security numbers, personal bank account and credit card numbers, and personal health information, unless this information is directly relevant to any employee's conduct relevant to the issues in the litigation.
- For the avoidance of doubt, information related to compensation, purchase of shares, or financial details relating to company acquisition are not encompassed within the definition of Confidential Personal Information or personnel files.
- Privileged information or attorney work product, as defined by relevant state law may be withheld so long as the metadata that would be present in a privilege log is provided.
- Trade secret material, as defined by relevant state law may be withheld for 3 years after the data of document creation, so long as enough metadata are made available to understand the topic of the document. Trade secret claims may be renewed for additional 3-year periods after review by the independent agent with the authority to resolve any disputes.

There is precedent for these provisions in the 2006 Final Judgment and Remedial Order (Order #1015) in *U.S. v. Philip Morris* (which requires defendants to review all trade secret assertions every three years to determine whether they still satisfy the definition of "trade secret")[5] and in the 2021 Judgment in *Commonwealth of Massachusetts v. McKinsey & Company* (which requires defendants to review all trade secret assertions after a period of seven years and to produce unredacted copies).[6]

Redactions should be completed by a company within 3 months of the settlement and a redaction log be created by that company and made public. A company should provide the corresponding metadata records for the withheld or redacted documents, giving users a complete picture of the entire corpus of documents.

If a company does not meet this deadline the documents should be provided to the document repository in partially redacted or unredacted form together with necessary funding so the repository can complete the redaction process.

The Attorney General or other plaintiff should retain unredacted forms of the documents so that future disputes can be resolved.

An independent authority to resolve disputes over redaction, privilege and trade secret issues should be identified. The defendant should pay the costs of maintaining this authority.

The UCSF Library or other repository should create a process for applying additional redactions if the need arises later.

[5] (U.S. v. Philip Morris USA, Inc., et al. Order #1015 Final Judgment and Remedial Order, 2006, p. 16)
[6] (Commonwealth of Massachusetts v. McKinsey & Company, Inc, United States. Assented-To Motion for Entry of Judgment, 2021, p. 12)

5. **A procedure should be established by which members of the public may challenge the appropriateness of a redaction or withheld document(s) and appeal to have that document(s) reviewed and released.**

Members of the public, including document repository staff, should have the ability to request that any document which has been redacted or withheld be reviewed and released by a company if the document does not, or no longer, contains information which must be protected under the provisions outlined in Recommendation 4 above. The 2011 consent order in *U.S. v. Philip Morris* established this procedure for the tobacco documents, which the UCSF Library, working with the US Department of Justice, helps to facilitate.[7]

6. **If a company does not provide metadata, PDF files, OCR text, or perform specified redactions in a timely manner, additional funding must be provided to enable a documents repository to do this work.**

Processing documents to create metadata, generate OCR text, create PDF access copies, and to identify and redact protected information incurs significant additional expense (detailed below) which should be reflected in any cost estimates.

7. **Funding to process and maintain long term free public access should be included in the settlement or judgement.**

Preserving and maintaining public access to digital materials in the long-term requires sustainable funding. Although some physical materials can theoretically exist in a state of "benign neglect" for years without great risk of loss, digital archives require active management to protect against file corruption ("bit rot"), hardware/ software obsolescence, and storage media failure, and to maintain a functional user interface and access point.[8]

A successful model has been used for more than two decades to support UCSF's Truth Tobacco Industry Documents, which in 2001 received $7.5 million (equivalent to $11.3 million in 2021 dollars; half the total funds to UCSF described above) from the American Legacy Foundation (ALF), which was created and funded by the Tobacco Master Settlement Agreement. This $7.5 included a $5 million endowment ($7.5 million in 2021 dollars) that has ensured the availability and longevity of public access to the tobacco documents at UCSF, which, in turn, enabled the development of a robust worldwide research community which has collectively produced over 1,000 scientific papers and reports citing the documents, leading to life-saving work in global tobacco control, public health policy, and ongoing tobacco litigation.[9]

---

[7] (Order #27 Remand: Consent Order Between the United States, the Public Health Intervenors, Philip Morris USA Inc., Altria Group, Inc., and R.J. Reynolds Tobacco Company Concerning Document Disclosure Obligations Under Order #1015, 2011)
[8] See (DeRidder, 2011), (Ovenden, 2019)
[9] (UCSF Industry Documents Library)

This funding model was sufficient to acquire, process, preserve, and maintain public access for the original 40 million pages of tobacco documents disclosed as a result of the Master Settlement and for ongoing document disclosures mandated through 2010. However, the final court order in *U.S. v. Philip Morris* required the tobacco companies to continue making their documents public for a period of fifteen years, which extended the MSA's original date for another eleven years until 2021. Over that period the IDL has acquired and preserved an additional 3.6 million documents which has put pressure on the original endowment. In 2011, the U.S. Department of Justice secured a consent order that provided $6.9 million ($8.5 million in 2021 dollars) from the tobacco companies through the court.[10] These funds were provided to UCSF improve public access and to enhance metadata.

The example of the Snowden Archive illustrates the difficulties of maintaining a digital repository over the long term without sustainable funding. The Intercept created the Snowden Archive to house the vast trove of National Security Agency documents leaked by Edward Snowden in 2013, but its parent company shut down the archive in 2019 citing "other editorial priorities" and encouraged the archive's creators to "find a new partner – such as an academic institution or research facility – that will continue to report on and publish the documents in the archive consistent with the public interest."[11]

Costs

The costs for preserving and providing long-term public access to millions of documents include: 1) initial costs of data servers and storage; 2) creation of OCR text if required; 3) redaction of protected information if required; 4) indexing (creation of metadata) if required; 5) trained personnel to actively monitor the files, provide user support, and maintain and update the technical infrastructure; and 6) long-term document storage and maintenance in perpetuity. As noted above, if the documents are not provided in digital form, there will be additional costs to digitize them.

Data Servers and Storage

The IDL currently uses Amazon Web Services (AWS) to store, back up, and serve data, as AWS has been identified as the most cost-effective option. The average cost for all functions related to document ingest, processing, storage, backup, and public access is $0.96 per GB per year. Cost estimates should account for the original data, plus processed data such as PDF access files, metadata records, extracted text files, thumbnail images, and backups of the original and processed data. We have found that the total storage required may be up to nine times the file size of the original data. The annual budget for data servers and storage in FY2020-2021 for 15 million documents (55 TB) was $55,000.

[10] (Order #27 Remand: Consent Order Between the United States, the Public Health Intervenors, Philip Morris USA Inc., Altria Group, Inc., and R.J. Reynolds Tobacco Company Concerning Document Disclosure Obligations Under Order #1015, 2011)

[11] (Society of American Archivists Human Rights Archives Section, 2019)

Creation of OCR Text
As described above, OCR text is required to screen documents to identify and redact any personal information, and to enable full-text search. It can be generated from TIFF or PDF files if it is not provided with the original data. The costs to generate OCR text include: data server(s) to process the files; use and maintenance of OCR software such as Amazon Textract, ABBYY FineReader, Tesseract, or iText (including software license and support fees); and staff costs to monitor and perform quality control checks on the OCR output. Depending on the extent and image quality of the documents, and on the type of software required, OCR costs may range from $0.0013 to $0.004 per page. For example, for 10 million pages (estimated 2.5 million documents) this is a cost of $13,000 to $40,000.

Redaction of Protected Information
Documents cannot be made available for public access if they contain legally-protected information. **We strongly recommend that documents be redacted prior to transfer to a public documents repository, as long as this can be completed in a timely manner and these is an efficient process for challenging company redactions.**

If documents are not redacted, there are significant additional costs involved in screening files to identify and locate all protected information and to apply and document appropriate redactions. Based on an estimate from a third-party de-identification vendor, these costs may range from $0.35 to $0.75 per page.[12] For example, a collection of 10 million pages could incur costs of $350,000 to $750,000 for screening and redaction.

Metadata and Indexing
Each document must be described with the minimal metadata fields listed in the Metadata Specifications above so that it is discoverable among millions of other documents. If metadata is missing it must be created manually, supported by automated methods. Previous costs incurred by the IDL for manual indexing range from $0.15 to $0.58 per page, depending on the number of metadata fields to be completed. A recent project to create detailed metadata for 207,824 pages at $0.52 per page cost $108,069.

Automated indexing using text analysis (including Natural Language Processing and Named Entity Recognition) and machine learning is becoming an increasingly viable and cost-effective solution. However, automated indexing is not yet reliable or scalable for documents containing handwriting, images, or with poor-quality extracted text.

Personnel
The IDL currently employs 4.15 FTE which includes archivists, software developers, and administrative staff. This team has the capacity to collect, process, and make public approximately 50,000 documents per month; provide reference services and other user support; perform regular software updates, security checks, user interface upgrades, and other technical maintenance; and conduct education and outreach activities to benefit current and potential archive users. The personnel budget in FY2020-2021 (including benefits) was $730,000.

[12] (Braided Data Solutions)

Long-Term Data Storage
Preserving the original and processed data, and maintaining a technical environment for public access, incurs ongoing costs. Although data storage costs are decreasing every year it is still a significant annual expense to store, backup, and provide public access to millions of documents. As noted above, the ongoing annual budget for data servers and storage is currently $55,000 for 15 million documents.

Endowment Funding Model
The tobacco documents archive has been successfully maintained for nearly 20 years thanks to a restricted $5 million endowment which generates sufficient income to cover annual data costs and essential personnel. For the reasons outlined above, future document disclosure initiatives should include an endowment to pay for long-term preservation and access to the documents.

Cost Scenarios

As an example, we estimate costs below for a collection of 2.5 million documents (10 million pages).

A) **In a best-case scenario**, where documents are in digital form and: 1) are redacted prior to transfer to a repository; 2) are produced in native, TIFF, and PDF format and accompanied by OCR text containing page-break indicators; 3) are indexed with full metadata; and 4) require little intervention by staff, the minimum annual cost for maintaining, preserving, and providing access to this collection is approximately $125,500 ($0.012 per page). An endowment of $2.9 million ($0.29 per page) would be needed to generate sufficient income to support this annual cost in perpetuity, bringing the total combined cost for upfront processing plus long term preservation and access to **$3 million ($0.30 per page)**.

B) **In a medium-case scenario**, where the documents are provided digitally but: 1) contain protected information and are unredacted; 2) are produced in native format only with no accompanying OCR text; 3) do not include sufficient metadata; and 4) require significant intervention and management by staff, the minimum upfront cost to process is approximately $2.4 million ($0.24 per page), followed by annual costs for preservation and access services of approximately $125,500 ($0.012 per page). An endowment of $2.9 million ($0.29 per page) would be required to generate sufficient income to support this annual cost in perpetuity. The combined cost of upfront processing and long-term preservation and access is **$5.3 million ($0.53 per page)**.

C) **The worst-case scenario** would be one in which the documents are produced on paper and require digitization. Estimated costs would include digitization (approximately $0.36 per page) and shipping, in addition to: creation of OCR text; creation of metadata; review and redaction as needed; processing by staff; and

long-term preservation and access. The cost to digitize 10 million pages is approximately $3.6 million, which combined with the costs listed in B) brings the total cost to **$8.9 million ($0.89 per page)**.

The UCSF Library is available to consult (at no cost) with Attorneys General and others negotiating settlements to develop specific cost estimates that reflect the realities of individual cases and settlements.

Additional Comments on Preservation of Chat Messages and Channels (Slack)

Production and preservation of chat messages from platforms such as Slack is an issue that is only just beginning to be investigated by the legal and archival professions. Slack offers various options for exporting data depending on the type of permissions and subscription held by the user.[13] The exports contain a workspace's message history in JavaScript Object Notation (JSON) format and include file links from all public channels. Every Slack message in a JSON file will include the following fields at minimum:

- type: indicating that the data is a message (or other type)
- user: the ID of the Slack user who sent the message
- text: contains the text of the message
- timestamp ("ts"): the time the message was posted (in Unix timestamp format)

Additional fields may be present if, for example, a message has attachments, was starred or pinned by a user, or received emoji reactions from other users. Edited messages may include a field showing the original unedited text. These and other fields are all detailed in the Slack guide on how to read messages exported in JSON files[14].

For organizational accounts, Slack provides access to its Discovery Application Programming Interface (API), which can integrate with eDiscovery and data loss prevention (DLP) solutions. Several eDiscovery companies offer software and services to interpret the JSON export in a more human-readable format.[15]

From an archival perspective, the JSON export is suitable for long-term preservation. The Library of Congress Recommended Formats Statement (RFS) includes JSON as a preferred format for datasets.[16] **Therefore, IDL recommends preserving the original JSON export in a documents repository**. If a JSON file is produced it should include all applicable fields from the eDiscovery Metadata Specifications above (including, but not limited to, date the JSON file was created, the JSON filename, file size). Metadata for each individual Slack message should also be included in the JSON file as noted above.

[13] (Slack, 2021a)
[14] (Slack, 2021b)
[15] For example: (Logikcull, 2021) and (Onna, 2021)
[16] (Library of Congress)

From an access perspective, the JSON export presents challenges because it is not easily readable to the average user unless the data is presented in an appropriate viewer. However there are various Slack export viewer tools available which could be adopted by a documents repository or by an individual user.[17] The Slack application itself can also be used to import the JSON data and recreate Slack messages and public channels.[18]

Conclusion

As Dr. Stanton Glantz wrote in a 2019 Op-Ed for *The Washington Post*, "lawsuits against companies aren't just about getting money. They're about revealing the truth."[19] Document disclosure is a powerful action by state attorneys general and others prosecuting cases against companies like Juul to pursue transparency, accountability, and justice. The groundbreaking effort for disclosure from the Tobacco Master Settlement Agreement enabled the creation of the Truth Tobacco Industry Documents Library and led to significant contributions to life-saving research and public health policies and laws. UCSF offers these technical recommendations for preserving industry documents in a cost-effective and sustainable model with the goal of supporting similar efforts to shine a light on industry actions, and to continue the drive to investigate these factors and protect public health.

For more information please contact:

Kate Tasker, Industry Documents Library Managing Archivist
kate.tasker@ucsf.edu
www.industrydocuments.ucsf.edu

[17] For example: (Faran, 2021), (JSONviewer, 2021), or (Backupery, 2021)
[18] (Slack, 2021c)
[19] (Glantz, 2019)

Works Cited

Backupery. (2021). *Backupery for Slack Export.* Retrieved July 2021, from Backupery: https://www.backupery.com/products/backupery-for-slack-export/

Braided Data Solutions. (n.d.). *Pricing.* Retrieved July 2021, from Braided Data Solutions: https://braideddata.com/pricing/

Commonwealth of Massachusetts v. McKinsey & Company, Inc, United States. Assented-To Motion for Entry of Judgment, 2184CV00258 (Suffolk Superior Court February 4, 2021).

DeRidder, J. L. (2011, June 1). Benign Neglect: Developing Life Rafts for Digital Content. *Information Technology and Libraries, 30*(2), 71-74.

Digital Preservation Coalition. (n.d.). *Digital Preservation*. Retrieved July 2021, from Digital Preservation Coalition: https://www.dpconline.org/digipres

Faran, H. (2021, February 24). *Slack Export Viewer 1.1.0.* Retrieved July 2021, from GitHub: https://github.com/hfaran/slack-export-viewer#readme

Fernandez, E. (2011, December 13). *UCSF to Receive Tobacco Papers, Funding to Improve Public Access to the Documents.* Retrieved July 2021, from UCSF: https://www.ucsf.edu/news/2011/12/98482/ucsf-receive-tobacco-papers-funding-improve-public-access-documents

Glantz, S. (2019, September 9). Opinion: Lawsuits against companies aren't just about getting money. They're about revealing the truth. *The Washington Post*: https://www.washingtonpost.com/opinions/2019/09/09/lawsuits-against-companies-arent-just-about-getting-money-theyre-about-revealing-truth/

JSONviewer. (2021). *Want to Easily View and Search Your Slack Workspace Export?* Retrieved July 2021, from JSONviewer: https://jsonviewer.co/

Library of Congress. (n.d.). *Digital Preservation at the Library of Congress*. Retrieved July 2021, from Library of Congress: https://www.loc.gov/preservation/digital/index.html

Library of Congress. (n.d.). *Recommended Formats Statement.* Retrieved July 2021, from Library of Congress: https://www.loc.gov/preservation/resources/rfs/data.html

Logikcull. (2021). *The Lawyer's Guide to Discovery and Investigations in Slack.* Retrieved July 2021, from Logikcull: https://www.logikcull.com/slack

Onna. (2021). *The Beginner's Guide to Slack eDiscovery.* Retrieved July 2021, from Onna: https://onna.com/blog/the-beginners-guide-to-slack-ediscovery/

Ovenden, R. (2019, Octobere 10). *Libraries' Role in Preserving Digital Information.* Retrieved July 2021, from Carnegie Reporter: https://www.carnegie.org/news/articles/libraries-role-in-preserving-digital-information/

Public Health Law Center. (n.d.). *United States v. Philip Morris (D.O.J. Lawsuit)*. Retrieved July 2021, from Public Health Law Center: https://www.publichealthlawcenter.org/topics/commercial-tobacco-control/commercial-tobacco-control-litigation/united-states-v-philip

Schaefer, S., Chodacki, J., Ismail, S., Janée, G., Lopatin, E., Macquarie, C., . . . Troy, S. (2020, August 10). *University of California Digital Preservation Strategy Working Group: Phase Two Report.* Retrieved July 2021, from University of California Libraries: https://libraries.universityofcalifornia.edu/doc/projects-and-groups

Slack. (2021a). *Guide to Slack Import and Export Tools.* Retrieved July 2021, from Slack Help Center: https://slack.com/intl/en-gb/help/articles/204897248-Guide-to-Slack-import-and-export-tools

Slack. (2021b). *How to Read Slack Data Exports.* Retrieved July 2021, from Slack Help Center: https://slack.com/intl/en-gb/help/articles/220556107-How-to-read-Slack-data-exports

Slack. (2021c). *Import Data from One Slack Workspace to Another.* Retrieved July 2021, from Slack Help Center: https://slack.com/help/articles/217872578-Import-data-from-one-Slack-workspace-to-another

Society of American Archivists Human Rights Archives Section. (2019, March 28). *The Intercept Shuts Down Access to Snowden Trove (Daily Beast).* Retrieved July 2021, from Society of American Archivists: https://www2.archivists.org/groups/human-rights-archives-section/the-intercept-shuts-down-access-to-snowden-trove-daily-beast

The U.S. National Archives and Records Administration. (2019, September 16). *Digital Preservation Strategy*. Retrieved July 2021, from National Archives: https://www.archives.gov/preservation/electronic-records/digital-preservation-strategy

U.S. v. Philip Morris USA, Inc., et al. Order #1015 Final Judgment and Remedial Order, Civil Action No. 99-2496 (GK) (United States District Court for the District of Columbia August 16, 2006).

U.S. v. Philip Morris USA, Inc., et al. Order #27 Remand: Consent Order Between the United States, the Public Health Intervenors, Philip Morris USA Inc., Altria Group, Inc., and R.J. Reynolds Tobacco Company Concerning Document Disclosure Obligations Under Order #1015, Civil Action No. 99-CV-2496 (GK) (United States District Court for the District of Columbia December 15, 2011).

UCSF Industry Documents Library. (n.d.). *Bibliography.* Retrieved July 2021, from UCSF Industry Documents Library: https://www.industrydocuments.ucsf.edu/biblio