Title

Digital Archives and Data Science: Building Programs and Partnerships for Health Sciences Research

Permalink
https://escholarship.org/uc/item/6hc248qk

ISBN
9781799897026

Authors
Tasker, Kate
Taketa, Rachel
Macquarie, Charles
et al.

Publication Date
2022-05-01

Peer reviewed

Chapter 7
Digital Archives and Data Science:
Building Programs and Partnerships for Health Sciences Research

Kate Tasker
University of California, San Francisco, USA

Rachel Taketa
University of California, San Francisco, USA

Charles Macquarie
University of California, San Francisco, USA

Ariel Deardorff
University of California, San Francisco, USA

ABSTRACT

This chapter describes work by the UCSF Industry Documents Library to develop resources, programs, and initiatives to support data science work with a diverse audience in the fields of health sciences, history of medicine, public health policy, and tobacco control. The Industry Documents Library (IDL) is a digital archive of over 15 million documents created by industries impacting public health, hosted by the University of California, San Francisco (UCSF) Library. The chapter describes the public health impact of industry documents research, highlights several examples of computational projects conducted by IDL scholars, outlines the IDL's developing plans for using data science techniques to assist with large-scale digital collection appraisal and metadata enhancement, and discusses how the IDL is expanding its collaborations with the UCSF Library's Data Science Initiative and Archives and Special Collections departments to further develop impactful data science programs across the university.

DOI: 10.4018/978-1-7998-9702-6.ch007

INTRODUCTION

When the UCSF Industry Documents Library first launched in 2002, it was one of the earliest digital libraries to offer access to its collections on the web. From the beginning, librarians, archivists, technologists, and the public health research community at the University of California, San Francisco (UCSF) grappled with how best to manage, index, search, and analyze millions of digitized documents which had been released from historic litigation against the tobacco industry. As the interrelated fields of information science, digital archives, and data science evolved, the Industry Documents Library (IDL) developed its own software tools and approaches to meet researchers' "big data" needs, at a time when few out-of-the-box solutions existed and none fit the unique requirements of a project which combined elements of digital archives, law libraries, and corporate accountability. For two decades the IDL focused on stewarding its data, building custom open-source tools for search and access, and supporting new public health research methods for studying and responding to the spread of tobacco-related diseases. During this period the IDL collection grew from a few thousand pages to a few million pages, and now contains more than 94 million pages in over 15 million documents. As the volume of available material continues to increase – the IDL anticipates adding more than 5 million new documents in the next year – the need to build partnerships and design collaborative strategies for managing data at scale is more imperative than ever. The expanding data science landscape now provides significant opportunities for the IDL to move beyond custom isolated solutions, and to embrace an approach which engages with an advancing library and data science knowledge base, innovative new ideas developed by other libraries and archives, and an expanding network of potential collaborators and partners.

This chapter begins with an overview of the UCSF Industry Documents Library, and highlights examples of how IDL scholars have applied data science methods to investigate specific questions about the tobacco industry's impact on public health. It describes the application of emerging computational tools for digital archival appraisal and description, and the potential for these tools to dramatically improve the IDL's capacity for curating its data. The chapter then describes the IDL's current and potential collaboration opportunities with the UCSF Library's Data Science Initiative (DSI), which serves as a campus hub for education and support in data science. This is followed by a description of another UCSF archival initiative which serves as a model for the IDL's data science engagement efforts: the UCSF Archives and Special Collections project "No More Silence: Opening the Data of the HIV/AIDS Epidemic using Natural Language Processing Techniques." This project offers rich instruction in how to prepare data for computational analysis, build a data science research community, develop training workshops, and connect with other campus partners. The chapter concludes with recommendations identified by the IDL, UCSF partners, and others in the library and archives profession, on how to encourage staff development and skill building, build cross-campus partnerships, provide opportunities for mutual learning through data science student internships, integrate data science tools into existing workflows, and plan for the long-term sustainability of data science activities.

BACKGROUND

The UCSF Industry Documents Library

The Industry Documents Library is a digital archive of documents created by industries which influence public health, hosted by the UCSF Library as a part of the Archives & Special Collections unit. Originally established in 2002 to house millions of documents publicly disclosed in litigation against the tobacco industry in the 1990s, the IDL has expanded to include documents from the opioid/drug, chemical, food, and fossil fuel industries to preserve open access to this information and to support research on the commercial influences on public health. The IDL is currently supported by four full-time staff: a managing archivist, a processing and reference archivist, and two software developers. Strategic leadership is provided by the University Librarian and the Associate University Librarian for Collections/University Archivist.

The Industry Documents Library had its beginnings in 1994, when a box of Brown & Williamson Tobacco Company documents was sent by an anonymous whistle-blower to UCSF's Professor Stanton Glantz, a well-known and outspoken tobacco control advocate. The documents proved to be a treasure trove of internal reports and communications revealing Brown & Williamson had recognized since the 1960s that nicotine was addictive and tobacco smoke harmful, contrary to the company's public claims of doubt and disbelief (Glantz, Slade, Bero, Hanauer, & Barnes, 1998).

Glantz and a small team of researchers indexed the materials by hand and deposited the set of over 1,000 documents in the Archives and Special Collections Department at the UCSF Library. In an effort to block public scrutiny, Brown & Williamson positioned private investigators at the Library to discourage patrons from viewing the documents. At one point the company sued the University of California Regents to have the materials removed from the Archives, but the Court ruled in favor of the public's "right to know" (Chandler & Storch, 2002). To reach a wider audience, the UCSF Library leveraged the technology of the time and disseminated the documents to researchers by CD-ROM and through the nascent UCSF Library website.

Millions More Documents Produced in Tobacco Litigation

In 1994, the Attorneys General of four States -- Mississippi, Minnesota, Florida, and Texas -- separately filed lawsuits against the tobacco industry for reimbursement of health care expenditures arising from tobacco-related illnesses. During the course of this litigation, the rest of the States joined in similar legal actions. In 1998, the state of Minnesota reached a settlement with the five major US tobacco companies, British American Tobacco, and two other tobacco-related organizations. One of the provisions of the settlement was the creation of two depositories into which the companies had to place the millions of documents produced in the case. The US companies' Minnesota Tobacco Document Depository and British American Tobacco's depository in Guildford, England were created and required to remain open to the public (UCSF Industry Documents Library, 2021).

A few months later in November 1998, 46 Attorneys General signed the Master Settlement Agreement (MSA) with the major tobacco companies in the US (California Office of the Attorney General, 2021). The MSA settled the remaining States' lawsuits by requiring yearly payments by the tobacco companies to the States and placing restrictions on advertising, marketing, and promotion of cigarettes. As part of the MSA, the companies were ordered to digitize and publish their internal documents produced for the cases on their own document websites as well as place physical copies in the Minnesota depository.

In 2000, the American Legacy Foundation (now known as the Truth Initiative), awarded funding to UCSF to create a permanent collection of the more than four million tobacco industry documents disclosed and digitized during the tobacco trials of the 1990s. The Legacy Tobacco Documents Library (now known as the IDL) was launched online on January 30, 2002.

In 2006, US District Judge Gladys Kessler ruled in a civil lawsuit brought by the Department of Justice that the nation's top tobacco companies violated the Racketeer Influenced and Corrupt Organizations (RICO) Act, misleading the public for years about the health hazards of smoking (United States v. Philip Morris USA, Inc., 2006). As a result of this case, the companies were obliged to make publicly available any documents produced for litigation on smoking and health until 2021.

For over 15 years, Library software developers scraped the tobacco industry documents websites monthly, adding new documents as they became available, updating metadata, and identifying and reporting any issues found in the documents. While the tobacco documents corpus was continuing to grow, UCSF staff responded to growing researcher interest in cross-industry practices and began to acquire documents from other industries as potential sources of study of commercial influences on health. The IDL now contains more than 94 million pages in over 15 million documents, which creates a landscape of complex data management challenges along with promising opportunities for the application of data science methods.

Scope and Impact of the Industry Documents Library

The collections are comprised of textual and graphic documents that have been converted to PDF; audio and video materials; metadata records from outside partner organizations; and a small set of proprietary file formats such as Excel or database files not viewable from within the Library but made available for download. Most of the collections have come to the Library in PDF format, whether as scanned copies of paper-based documents or born digital. The tobacco industry documents were transferred to UCSF already described with item-level metadata, but a majority of the other industry collections came without any indexing, necessitating the creation of metadata either in-house or outsourced to a vendor for larger collections.

To process a new collection, the IDL team takes the PDF documents and runs them through an Optical Character Recognition (OCR) process, creating TIFF images and full-text searchable PDFs. These searchable PDF documents are then matched to their respective records in the item-level metadata file. From the beginning, the Library focused on open-source tools and building in-house expertise. The first iteration of IDL, the Legacy Tobacco Documents Library (LTDL), used the University of Michigan's Digital Library Extension Service (DLXS), Apache, and a PostgreSQL relational database. A detailed history of the LTDL's technological development has been documented by Schmidt, Butter, & Rider (2002). Current software includes a MariaDB database and Solr search index, which underlies the online user interface which was built using a customized WordPress template. In addition to shareable open-source tools and software, the IDL is committed to fostering an environment of partnership, sharing metadata records across institutions like the Rutgers School of Public Health's Trinkets & Trash database and the Medical Heritage Library. The idea is the more researchers have access, the greater the investigations and resulting impact for public health.

Since 1994, researchers, lawyers, policymakers, and journalists have used the documents to shed light on the tobacco industry's harmful practices and its efforts to cover up these harms. Their work has resulted in the publication of over 1,100 peer-reviewed papers, government reports and media stories, which in turn have informed U.S. tobacco control legislation and a World Health Organization international treaty. The industry documents research community has also trained postdoctoral fellows and students who are impacting tobacco control policy and public health practice around the world.

NEW WAYS OF RESEARCHING WITH IDL

Computational Research Using the Industry Documents Library

Initially, the documents were accessible only through metadata fields (such as title, author, document date) and not through full-text search. Each tobacco company employed a slightly different metadata schema for their documents which created a host of data standardization problems and complicated searches across company collections. Despite these access barriers, early tobacco documents researchers methodically searched through and read millions of internal corporate documents between 1994 and 2004, producing approximately 300 published works which served as the foundation for deep investigations into the strategies and tactics of the tobacco industry.

In 2003, the UCSF Library undertook a very large project to create an OCR computer "farm." This project leveraged the processing power of idle computers in the building during off hours to run optical character recognition software to generate machine-readable text files for millions of documents. By 2004, all text-based documents in the Library had been processed in this way, resulting in full-text searchable PDFs. The ability to search the extracted text of a document, as well as its metadata, resulted in a flood of scholarly investigations, more than doubling research output to 700 additional papers, publications, and reports over the next decade. Optical character recognition (OCR) freed the text of every document for use in new and novel methods of research. With the emergence of computational research and digital humanities work, the IDL has seen documents research expand from the traditional and time-consuming document-by-document search and review technique, to using extracted text to create visualizations and trend analysis at the 10,000-foot view, glimpsing the forest as well as the trees.

New Ways to Investigate Industry Documents

The IDL team identified the need for more current data science tools a decade ago and responded by making it easier to programmatically query the Library's Solr server, where a copy of the entire index is stored. An Application Programming Interface (API) was created which allows users to easily export document records to another system, execute search queries, and process search results by program. Data can be exported in xml, json, python, ruby, php, and csv. In addition to the API, the IDL provides access to document text and metadata through downloadable datasets for each collection which are updated every year. If a user would like to work with the documents in their own database system, the IDL provides a MySql "database dump" file with the data and OCR text files for every text-based document.

Once researchers had the tools to download large batches of documents and their associated data, they were able to demonstrate new ways of looking at the documents to answer important research questions. Some key projects and methods include:

The Tobacco-Documents Project at the University of Georgia

The University of Georgia led an NIH-NCI grant-funded project from 2001 to 2004 using forensic and computational linguistics in the investigation of the tobacco industry's use of deceptive language within the tobacco documents corpus. The resultant Tobacco-Documents Project produced over 20 papers and publications as well as a publicly available text analysis toolkit (University of Georgia, 2005).

Proctor and Risi: Tobacco Analytics (Stanford University)

The Tobacco Analytics project began in 2015 and was led by Robert Proctor, Professor of History of Science at Stanford University, and Stephan Risi, then a PhD candidate in History, with support from the California Tobacco-Related Disease Research Program (TRDRP) High Impact Award and the Roberta Bowman Denning Fund for the Humanities and Technology at Stanford University. With IDL staff help, Risi acquired a data set of 11,303,161 tobacco documents and built the website tobacco-analytics.org, which featured visualizations of the dataset including frequency charts, ngram charts, and maps. The project was created to teach users new ways of researching in the tobacco documents corpus. The website provides case studies that use the project's tools and datasets to answer specific questions in a quantitative way, such as "When did smoking become an addiction?" (Proctor & Risi, n.d.)

Stephan Risi: Tobacco Networks (MIT)

The Tobacco Networks project was led by Stephan Risi and a team of students while Risi was an MIT postdoctoral associate in 2019. The project goal was to produce network graphs showing relationships and connections amongst researchers, lawyers, and marketing experts in the tobacco industry. These visualizations were crafted around specific research questions to serve as jumping off points for users who had no previous experience with the content of the tobacco documents. One research question they hoped to answer was, "How did the industry's research directors deceive the American public about the health harms of smoking?" (Risi, n.d.) To answer this particular question using traditional research methods, a researcher would first need to know the names of the research directors, who was the most connected, or who had the most dealings with outside entities. For a person new to the subject and content of the documents, the Tobacco Network project's visualization of key players could be vital as a starting point. Risi's team extracted data about the authors and recipients of documents from the 1970s, using letters, reports, and memos that were sent and received by persons associated with the tobacco industry. Students then computed a force-directed network graph and visualized the connections between key researchers at the various companies. The resultant network graph was implemented as an interactive web application which enables an 'at-a-glance' identification of the most highly connected research directors. A researcher unfamiliar with the content of this massive archive now has names of key people and some starting points from which to start their searches.

Risi and Proctor: Courtroom Tropes and Taboos

In this study Risi and Proctor worked with 318 closing argument transcripts from over 100 Engle progeny tobacco lawsuits, housed in the IDL's DATTA (Depositions and Trial Transcripts) collection. The authors used computational linguistics to identify differences in the rhetorical strategies used by lawyers in tobacco litigation. The team calculated frequency scores and Mann-Whitney Rho scores of plaintiffs versus defendants to reveal the "tropes" and "taboos" found in their statements to the juries.

Historically, the defendant tobacco companies have attempted to shift the focus of "blame" to the smoker by using terms like "free choice," and "personal responsibility." The use of data analysis revealed to Risi and Proctor how these tactics and strategies were used in current litigation without the lawyers actually speaking these expressions or phrases. Industry attorneys rarely mention personal responsibility, for example, but invoke that concept indirectly, by talking about "decisions" made by the individual smoker and "risks" they assume (Risi & Proctor, 2020). Here quantitative analysis reveals what could have been hidden patterns in courtroom rhetoric, including the avoidance of certain terms (taboos), such as "profits" or "customer." While cigarette companies use words that focus on the individual smoker, plaintiff attorneys tend to use terms that refocus responsibility onto the industry.

Mathew, Karatzas, and Jawahar: Document Visual Question Answering

Document Analysis and Recognition (DAR) involves extracting information from images and converting it into a form that a machine can read, like OCR, table extraction, or key-value pair extraction. A 2021 case study by Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar introduced Document Visual Question Answering (DocVQA), a process by which an "intelligent" reading system is expected to respond to impromptu requests for information, conveyed through natural language questions, by human users. To do this, the algorithm needs to extract and interpret not only the textual context of the document images (handwritten, typewritten or printed) but also utilize visual cues like the layout of the document (page structure, forms, tables) or the style elements (fonts).

The DocVQA dataset consisted of 50,000 questions that were defined utilizing over 12,000 document images downloaded from across IDL's five major industry collections - tobacco, food, drug, fossil fuel and chemical. Questions and answers on the selected document images were collected through a web-based annotation tool. The annotation process was organized into three iterative stages with different users defining question and answer pairs for each image. This process ultimately produced 50,000 questions framed on 12,767 images. The authors shared this data set to encourage a "purpose-driven" approach in document image analysis and recognition research (Mathew, Karatzas, & Jawahar, 2021).

Using Data Science for Digital Archives Workflows at Scale

In addition to opening new avenues of research inquiry, the application of data science methodologies to the IDL data set offers new capacities for archival workflows. The IDL operates under some unusual conditions which often make it difficult to apply traditional archival practices. Two of the more challenging circumstances include rapid appraisal of very large volumes of digital records, and enhancement of externally created document-level metadata.

Data Science Tools for Digital Appraisal

As is widely discussed in digital archives literature (Anderson, Eaton, & Schwartz, 2015; Lee, 2018; Shallcross & Prom, 2016), the traditional practices for appraising physical archival materials must be updated and augmented with new approaches for digital environments. Where before an archivist may have visually inspected the contents of a file cabinet or office - noting physical condition, extent, types, and value of documents, skimming folder titles, and developing a general insight about the contents of the collection - the same practices are not feasible for very large digital collections of the type that the

IDL regularly accessions. Although the fundamental principles of archival appraisal remain the same, a new set of tools and workflows must be applied in the digital landscape. Archivists cannot tell from visually inspecting a hard drive how many folders and files it contains, their condition, or what the contents are generally about. Digital archive appraisal requires the use of software (ranging from basic desktop applications to powerful digital forensics tools) and the application of computational processes to generate file directory lists, identify corrupt or unreadable files, calculate file and folder size, and glean a sense of the contents of the collection.

At the IDL, digital collections do not arrive on their original source media. The files are digital copies which have been exported from legal ediscovery software, downloaded from a file-sharing application in the cloud, or sent electronically using file transfer protocol (FTP). The files originated with hundreds or thousands of individual creators but are usually contributed by someone who did not create the original copies (such as a plaintiff's lawyer, a journalist, or a whistle-blower). The context of creation is more difficult to determine as the appraising archivist has little information about the file system or overall folder structure. In the case of litigation materials, documents are exported from an ediscovery platform as a "production," based on specific criteria and search terms. A single file from a larger folder may be included in the production, with other folder contents excluded, or all files in the folder may be produced. Documents obtained from Freedom of Information Act (FOIA) requests are also selected using criteria specified in the request, and often are combined by the responding organization into a PDF binder or portfolio package which is a further step away from a document's original file format and location.

Given this absence of context, the IDL must rely on the information in the documents themselves to understand the content of the collection. Human review of millions of files is not possible and sampling the files can only offer limited insight given the lack of original order and file structure. In the case of the tobacco documents, the IDL is fortunate to have a long-standing and productive partnership with UCSF's Center for Tobacco Control Research and Education (CTCRE) whose director, 75 affiliated faculty members, and cohort of postdoctoral research fellows devote significant time to conducting research in the documents to advance their scholarship and to disseminate findings to policymakers and to the public health community. Over the past twenty years, CTCRE faculty and fellows have obtained dozens of grants to undertake major projects (often spanning years or decades) to identify and analyze the contents of the tobacco collections, resulting in an extensive body of published knowledge which provides context and insights for further research.

For other industry documents collections where this kind of robust research partnership is less established, or non-existent, the IDL and its collaborators have yet to secure the resources to grow and sustain a library and research ecosystem like the model that exists with the CTCRE. Although the IDL has active and rewarding collaborations with other UCSF centers (including the Philip R. Lee Institute for Health Policy Studies, the Program on Reproductive Health and the Environment, and the Environmental Research and Transformation for Health Center), the application of traditional research methods will still take many years and major grant funding to establish the kind of knowledge base which has been developed for the tobacco collections.

A promising solution lies with the collaborative application of data science methods, in particular natural language processing (NLP) approaches including topic modeling and named entity recognition (NER). In the archives field, the work to develop these methods for use with born-digital archival collections has been significantly advanced by the Mellon Foundation-funded BitCurator-NLP project which ran from 2016-2018 and built on previous work by the BitCurator project (2011-2014) and BitCurator Access project (2014-2016). Members of the BitCurator-NLP team developed software "to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections" (BitCurator, n.d.) The BitCurator project investigated Python-based tools gensim and pyLDAvis to generate and visualize topic models. Topic modeling aims to automatically identify the main topics found in a document or set of documents by clustering words based on various patterns and frequency. Topic modeling could help archivists more quickly grasp the general scope and content of a very large collection of documents, or pinpoint significant individual documents. The application of topic modeling is still in an early stage of development, with the BitCurator tools able to analyze a corpus as a whole but not drill further down into individual documents. An earlier tool called ArchExtract, developed at the University of California, Berkeley, included the ability to examine topics for individual documents, but unfortunately this tool is no longer in development (Elings, 2016). Although innovative and effective in specific use cases, the IDL has not had the resources or technical expertise available to thoroughly explore these tools. It appears that similar work at other institutions is mostly at an experimental or project-based implementations (Hutchinson, 2020) although communities of practice such as the BitCurator Consortium are working to address the gaps identified in training, collaboration, research, software development, documentation, integration, and code, and to encourage greater use of these transformative tools.

Data Science Tools for Archival Description

Archival description is another aspect of the IDL's work which would greatly benefit from the application of data science methods. The IDL has historically received documents accompanied by item-level descriptive metadata which has been produced by creating organizations or by law firms during litigation. The quality of this item-level metadata is extremely varied and can range from highly detailed profiles with more than sixty descriptive fields, to sparse or empty records simply labeled "No Title." The robust metadata available for some documents is a highly valuable research aid and enables complex and powerful search queries. However, the presence of high-quality metadata for some documents also highlights the inadequate metadata for other documents and creates unequal description and search results across the collections. Manual metadata clean-up and enhancement of more than 15 million records is far too time-intensive and cost-prohibitive to consider. The IDL has periodically investigated the use of NLP techniques to enhance item-level metadata over the past ten years, but the scale and variable quality of existing metadata and OCR text, and the lack of suitable tools, presented a high barrier to progress.

In the summer of 2021, the IDL hosted a pilot data science internship to refresh its exploration of NLP tools for metadata enhancement. The team worked with an undergraduate computer science student who initiated an independent project to test and evaluate a subset of NLP tools for named entity recognition (NER), specifically to identify and extract named geographic entities (such as cities, counties, states, and countries) from IDL datasets. The tools assessed included Natural Language Toolkit (NLTK), spaCy, Stanford NER, and Amazon Comprehend. This evaluation significantly updated and advanced the IDL's understanding of the current NLP landscape and produced actionable recommendations for applying these tools. The IDL plans to build on this successful pilot by providing further opportunities for students and early career data scientists to develop their skills by working to computationally enhance IDL collection metadata.

The opportunities presented by NLP tools are significant and encouraging, but the IDL's collections still present a challenging use case for currently available platforms, given the volume and heterogeneity of documents and file formats. Much work is required to develop methods to implement these tools, integrate them with custom-built command-line programming scripts and systems, and operationalize workflows which more formally align the IDL with established guidance and best practices, such as the FAIR Principles (making data Findable, Accessible, Interoperable, and Reusable) and CARE Principles (Collective benefit, Authority to control, Responsibility, and Ethics) for Indigenous Data Governance (Carroll, Herczog, Hudson, Russell, & Stall, 2021). As data science becomes more of a strategic priority for the IDL, the team has increasingly sought to partner with others in the larger data science ecosystem at UCSF, including the Data Science Initiative and Archives & Special Collections, to develop the knowledge and capacity to successfully investigate and apply these methods.

COLLABORATIONS IN DATA SCIENCE

Working with the UCSF Library Data Science Initiative

The UCSF Library's Data Science Initiative (DSI) serves as a library and campus hub for education and support for biomedical data science. The team's mission is to build computational and data skills in the UCSF community by providing education and resources to trainees, faculty, and staff. These goals are accomplished through a variety of services including hands-on workshops, community meetups, hacky hours, and expert consultations. The team's services focus on five main areas of support: research data management, reproducibility & open science, programming, bioinformatics, and statistics, with classes and workshops that range from introductory (Python 101) to more advanced (Python for Natural Language Processing). DSI teaching introduces learners to the FAIR Principles, emphasizing the importance of improving the Findability, Accessibility, Interoperability, and Reuse of data. From the earliest days the work of the Data Science Initiative has relied on extensive partnerships and collaborations. One of its earliest collaborators was the Bakar Institute, UCSF's computational health science group that is often seen as the nexus for data science research on campus. DSI members were also crucial in founding the campus' Natural Language Processing (NLP) group, a researcher-driven community that regularly hosts talks, workshops, and hackathons. Finally, the team has sought to develop librarian's computational skills through their involvement in the non-profit training organization The Carpentries, and specifically Library Carpentry, coordinating hands-on workshops that teach data and software skills to information professionals.

Over the years IDL team members have collaborated with the UCSF Library's Data Science Initiative in a number of informal ways. In order to enhance their own computational skills, IDL team members attended several DSI-hosted programming workshops and Library Carpentry workshops and have often reached out for advice and assistance. Recently, the teams met to brainstorm opportunities for more formal collaborations. As an outcome of this meeting, the DSI team has invited the IDL team to present at an upcoming NLP community meetup. This will be an opportunity to showcase IDL collections to interested researchers, and forge stronger connections with the NLP community. DSI team members are also looking for opportunities to incorporate IDL documents into the workshops that they teach, especially their introductory NLP workshops. The teams have also discussed developing new workshops that would introduce UCSF tobacco scholars to new programmatic methods for accessing IDL documents, whether through NLP approaches or querying the collection via the API. Finally, the teams are investigating mechanisms for co-hosting summer interns, so that undergraduate and graduate students can receive mentoring from the DSI team while working on directed data science projects that enhance IDL metadata and collections.

Learning from UCSF Archives & Special Collections: The "No More Silence" Model

The UCSF Archives and Special Collections department is a dynamic health sciences research center that contributes to innovative scholarship, actively engages users through educational activities, preserves past knowledge, enables collaborative research experiences to address contemporary challenges, and translates scientific research into patient care. UCSF Archives' mission is to identify, collect, organize, interpret, and maintain rare and unique material to support research and teaching of the health sciences and medical humanities and to preserve institutional memory. The department created the Tobacco Control Archives in 1994 (as a complementary effort to the Brown & Williamson Tobacco Documents project) to document the tobacco control movement in California, the people and organizations involved, and related campaigns and legislation. The Head of UCSF Archives has also provided leadership to the IDL since its formation. UCSF Archives houses a rich collection of manuscripts and papers of individuals, university archives, historical records of UCSF hospitals, and administrative records of regional health institutions, and holds preeminent collections documenting the history of the HIV/AIDS epidemic and the community's response. Recent grants from the National Historical Publications and Records Commission (NHPRC) and the National Endowment for the Humanities (NEH) have enabled the large-scale digitization of HIV/AIDS materials in partnership with other San Francisco Bay Area institutions, opening up new ways of working with data.

This digitization work enabled the development of the project "No More Silence: Opening the Data of the HIV/AIDS Epidemic," which provides direct computational access to the historical documentation of the San Francisco Bay Area's response to the HIV/AIDS epidemic. UCSF Archives created this project to facilitate work in the areas of digital humanities, computational cultural heritage and health sciences, and machine learning, and enable broad research and educational inquiries into the history of the epidemic by allowing researchers, students, scholars, and the public to ask questions across a large quantity and variety of historic records at once.

This work is important to bridge gaps between patient care, the lived experiences of people with AIDS, and the historical and cultural components of the epidemic. The datasets that cover the quantifiable biomedical and statistical components of AIDS/HIV are many and massive, but this work attempts to address the lived experience of AIDS/HIV. The cultural process of living with AIDS, making meaning from it, and politicizing that experience is contained in many archives, but in few datasets. This project aims to address these questions, allowing researchers, communities, and publics to ask questions of a large corpus of records at once and trace such nuanced things as the adoption of diverse terminology around gender across time and institutions, the sentiments around different treatments and therapies as they entered the arena of possibility for the first time, and the significance of digital technologies in the broadening community response to AIDS & HIV.

There are three areas of work that the project team creates and maintains to support this goal:

- **Data:** create and maintain a dataset of textual and image data representing nearly 200,000 pages of archival documents and their associated archival metadata.

- **Computer Code:** create and provide access to scripts, computer programs, and code as examples and demos for accessing and using the data contained in the dataset.
- **Educational Workshops:** teach workshops using the dataset as a foundation to build beginning computer-programming literacy for researchers, scholars, and members of the public wanting to work in the areas of digital humanities and health sciences.

Developing Data Science Expertise

One of the most important facets of this project -- and not one that is usually touted in the public-facing documentation -- is the way in which it necessitates a collaborative approach across UCSF Library departments, UCSF on the whole, the University of California (UC), and even other institutions and individuals outside of the UC and the United States. The project team included Archives staff as well as a software developer from the Industry Documents Library. This combined the skillset and subject knowledge of archivists with the programming expertise of the IDL developer and encouraged greater exchange between Archives and the IDL. The team also worked to develop their own data science skills throughout the course of this project and partnered with the Library's Data Science Initiative to scope and vision a project which would be sustainable and would have use for data science practitioners already working in the field. This collaboration led to subsequent partnerships and networks that ran across research areas, such as with the UCSF NLP group -- a campus research initiative that Archives staff learned of through initial conversations with UCSF DSI team members. The IDL team received frequent updates on the project and was invited to learn and participate in "No More Silence" workshops and outreach.

At the outset, it was also clear that the use of the dataset created as a part of this project actually necessitated a skill set -- computer programming and computational analysis -- that was not widely present in Archives' existing user communities. Teaching users the skills necessary to use this data resource were a central component of the project, and while some Archives staff members had been learning some of these skills, much of this expertise needed to come from outside the Archives. Through largely fortuitous circumstances the project team was introduced to Dr. Clair Kronk, a PhD researcher working at the intersections of biomedical statistics and gender, sex, and sexual orientation and the creator of the Gender, Sex, and Sexual Orientation Ontology (GSSO) (Kronk & Dexheimer, 2020). The project team partnered with Dr. Kronk to develop and present curriculum for workshops on introductory computational text analysis and natural language processing using the No More Silence data.

Creating and Packaging a Historical Dataset

The UCSF Archives has, since the late 1980s, been collecting, processing, preserving, and providing access to archival materials that document the community response to the AIDS/HIV epidemic through the AIDS History Project (UCSF Archives & Special Collections, 2021). Recently, the Archives have been able to digitize large quantities of these materials through two grants from the National Historic Records and Publications Commission (Ilieva, 2017) and the National Endowment for the Humanities (Amin, 2017). The immense labor of digitization and metadata creation undertaken through these projects was key to allowing these materials to be relatively easily re-packaged as a dataset through the No More Silence project. With the materials already digitized and processed with OCR, the project team was able to work with the California Digital Library, which implements and maintains the Digital Asset Management Software used by Archives to store, describe and provide access to its collections, to harvest digital files, text, and metadata from all AIDS History Project collections and package them into the dataset. It should be noted that these are all materials that were already publicly available online to researchers, but this dataset creation and packaging step was what allowed these materials to be accessible all at once for computation research. Given the sensitive nature of some documents, access to the dataset is currently mediated by Archives staff. Interested researchers are encouraged to contact Archives staff for more information.

Partnering with other Campus Units

As mentioned before, forming partnerships with other UCSF campus units has been one of the most valuable parts of this project. One of the most notable has been the partnership with the UCSF NLP group. Through connections initially made with the help of the UCSF Library DSI, Archives and UCSF NLP put on a collaborative event called "Summer of NLP," where NLP practitioners and group members used the data in the No More Silence dataset as source materials for summer-long learning projects intended to build skills and demonstrate possibilities in NLP computer programming. The No More Silence project team introduced participants to the data; participants formed teams around intended subject/project areas and created self-directed work and scope plans for completing their projects. Projects were presented in a public event at the end of the summer. Teams completed fascinating work, including building tools to visualize and compare connotations of words throughout time across the epidemic and tracking the emergence of AIDS "topics" groupings of linked concepts frequently appearing together (Gologorskaya, Kaplan, Cahn, Mills, & Panahiazar, 2019), and mapping and comparing "hotspots" of AIDS/HIV activism and new AIDS case counts (Thombley, Renslo, Tonner, & Gopez, 2019). The IDL software developer, as a member of the project team and Summer of NLP workshop instructor, was also able to build relationships with the DSI and NLP communities, which has led to productive discussions of how a similar research opportunity could be developed using the IDL data.

Developing Community Outreach and Education Workshops

Though it was not always planned as such from the outset, outreach and the presentation of educational workshops and classes has become the most important aspect of the No More Silence project. Based on UCSF Archives existing knowledge of its users and patrons, the project team had some idea that skill-building in Archives' user communities would likely be needed if this data resource was to be used in the intended manner. As the project work commenced, it quickly became clear that the skillsets required to effectively use this resource -- namely computer programming and computational text analysis -- were almost non-existent among UCSF Archives users, and that building these skillsets would be an essential part of the project scope if the project were to be successful.

In addition, this work was not just the development of skills within a community but was also the development of a community of practice itself. As much as the workshops functioned to build skills, they also functioned to gather together practitioners around their interest in this work and forge connections -- both to each other and to resources -- that could help people in their work moving forward.

Materially, this happened through two processes: 1) "skilling up" on behalf of the project team, whose members took classes and pursued other online learning to build skills in teaching and writing computer code, especially in the programming language Python and the areas of Natural Language Processing and Machine Learning, and 2) the forging of partnerships as core components of the project, especially in the later stages.

Of special note is the forging of partnerships. The project benefited immensely from the contributions of Dr. Clair Kronk, who assisted in the development of workshop curriculum and who co-taught workshops with the project team. Additionally, through existing institutional partnerships with the Gay, Lesbian, Bisexual, and Transgender (GLBT) Historical Society, the project team was introduced to Krü Maekdo and the Black Lesbian Archives which she founded (Maekdo, 2019). Krü's contributions to community engagement around this data resource were invaluable, and the success of many of the project workshops from a community standpoint were largely due to her.

Impact on the Industry Documents Library

The No More Silence project provided a fantastic opportunity for the IDL to observe and learn how to develop an impactful community-centered data science program using archival collections. IDL team members heard about the triumphs and challenges of the project during regular Archives team meetings, attended NLP workshops, and, in the case of one IDL developer, worked directly with the data and with researchers as a member of the project team. This participation offered valuable insight into practical steps the IDL can take to implement data science workshops and trainings for IDL staff and researchers, develop an IDL data science community, and maximize the computational potential of the IDL collections.

FUTURE DIRECTIONS

Supporting the use of digital resources for academic research by developing and nurturing a data science ecosystem is clearly emerging as a key strategic priority for the IDL, campus departments, UCSF, the University of California, and the academic library profession as a whole. The UC Systemwide Libraries Annual Plans & Priorities document for fiscal year 2017-2018 recognized "long-term access to digital content" as one of its ongoing projects (University of California Libraries, 2017), a decision which was highlighted by the archivist and librarian data science community in the final report of the innovative Always Already Computational: Collections as Data project (Padilla, et al., 2019). The Always Already Computational project, led by Thomas Padilla from 2016-2018 with funding from the Institute of Museum and Library Services, brought a community of cultural heritage practitioners together to generate discussion, best practices, and useful guides for institutions as they prepare collections for use in computational research. Always Already Computational was followed by Collections as Data: Part to Whole (2018-2021) which is supporting project teams of cultural heritage professionals and disciplinary scholars to develop models for the creation and use of collections as data. Other libraries in the University of California system are continuing to develop their data science programs, such as the UCLA Library's Data Science Center, the UC Davis Library DataLab, and the UC Berkeley Library Data Initiatives Plan. A similar strategic focus has been applied at other academic libraries to further develop a framework for building successful data science initiatives (Mani, Cawley, Henley, Triumph, & Williams, 2021).

This catalyzing work provides clear recommendations which the IDL and other organizations can use to further guide its data science efforts. First, the need to support staff to develop data science knowledge and skills, by offering additional training and by encouraging the creative integration of data science methods and tools into existing projects. IDL staff have begun applying knowledge from Library Carpentry workshops by using OpenRefine, a powerful open-source tool for cleaning and analyzing large datasets, for some metadata projects, and intend to take advantage of further training opportunities offered through the UCSF Library's Data Science Initiative. Providing additional training and support for IDL staff is a foundational step which must be taken in order to better understand and support researchers' data science goals, and to thoughtfully and responsibly undertake these new initiatives to align with criteria like the FAIR and CARE Principles, and other important ethical considerations.

Second, the importance of taking an interdisciplinary and cross-campus approach to identify and nurture data science partnerships. Through increased collaboration with the DSI, UCSF Archives, and other partners, the IDL will look to expand outreach efforts to communicate with computational researchers at UCSF through webinars, workshops, blog posts, and social media interaction to better understand and meet user needs for finding and working with IDL data, and to expand existing services and tools such as IDL's API and prepared datasets. The IDL also has an opportunity to play a vital role in strengthening connections between interdisciplinary researchers who use IDL data by serving as a "hub" or place for information exchange and interaction.

Third, the value of creating opportunities for mutual learning by providing support for data science interns, fellows, and researchers to develop their skills at the library, with the added benefit of further improving IDL collections for data science projects. With the DSI, the IDL is planning a pilot initiative to co-host a summer data science fellowship which will provide a network of data science experts, librarians, and archivists as mentors for fellows conducting data science projects using IDL data.

Fourth, the importance of integrating data science opportunities into existing workflows, by making informed decisions about how the IDL can acquire, process, and manage its data to make it easier to implement available computational tools and enhance its collections as data. For example, data science is a key consideration in the planning and development of the Opioid Industry Documents Archive, a significant new collaboration between UCSF and Johns Hopkins University which builds on the tobacco industry documents model to preserve and provide public access to millions of documents being disclosed from U.S. opioid litigation. As this project develops, building in workflows for data science and roles for data science experts and advisors will be critical in effectively stewarding these documents and maximizing their impact for public health.

Fifth, and finally, the need to plan for the responsible and long-term sustainability of these new initiatives. Ames and Lewis argue that "Big Data in the library emphasises existing problems and brings about new questions of ethics, methodologies and transparency" and "requires organisation-wide re-evaluation" (Ames & Lewis, 2020). Padilla urges libraries to explore data science, machine learning, and artificial intelligence using the principle of "responsible operations" and to consider individual, organizational, and community capacities for implementing new technologies in ways that do not perpetuate harm (Padilla, 2019). As has been observed in reports of early data science projects and software development at academic libraries, unless there is permanent staffing and funding this work is often limited to one-off projects which are eventually paused or abandoned when resources run out (Hutchinson, 2020). These difficulties are not unique to data science efforts, and the academic library community is accustomed to the challenge of having to gradually build and evaluate programs through pilot projects, temporary staffing, and grant funding. Encouragingly, the work accomplished in these projects also contributes to the foundational knowledge, exploration, and collective advancement of long-term solutions. However, this precarity emphasizes the need for developing strong partnerships with collaborators who share the same goals and values for the advancement of ethical data science in academic libraries and archives, to maximize the investment of available resources and to implement meaningful and useful tools and services. These considerations will be critical areas of focus as the IDL continues to engage with data science work.

CONCLUSION

The exploration of data science methods and services in academic libraries and archives offers tremendous opportunities, as evidenced by the surging interest, discussions, and new initiatives proliferating across the profession. From the IDL's perspective as a twenty-year-old digital repository which has historically and often of necessity operated at the edges of more traditional archival practice, it is exciting to learn from the growing communities of library and archives practitioners, public researchers, and other scholars contributing their substantial expertise to the application of data science to library collections. As this innovative new landscape emerges, it will be essential to explore data science opportunities, and their challenges, by continuing to embrace partnerships, seek community, and create spaces for mutual learning.

ACKNOWLEDGMENT

This research was supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian [grant numbers 40-8869 and 40-8963]; and in part by California Revealed as made possible by funding from the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian [grant number 40-9043].

REFERENCES

Ames, S., & Lewis, S. (2020). Disrupting the library: Digital scholarship and big data at the National Library of Scotland. *Big Data & Society*, *7*(2). Advance online publication. doi:10.1177/2053951720970576

Amin, K. (2017, April 17). *UCSF Archives $315,000 to digitize AIDS archives.* Retrieved from UCSF Library: https://www.library.ucsf.edu/news/neh-awards-ucsf-archives-315000-to-digitize-aids-archives/

Anderson, B., Eaton, F., & Schwartz, S. (2015). Archival Appraisal and the Digital Record: Applying Past tradition for future practice. *New Review of Information Networking*, *20*(1-2), 3–15. doi:10.1080/13614576.2015.1114823

BitCurator. (n.d.). *BitCurator NLP*. Retrieved from BitCurator: https://bitcurator.net/bitcurator-nlp/

California Office of the Attorney General. (2021, November 24). *Master settlement agreement*. Retrieved from State of California Department of Justice, Office of the Attorney General: https://oag.ca.gov/tobacco/msa

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(108), 108. Advance online publication. doi:10.103841597-021-00892-0 PMID:33863927

Chandler, R. L., & Storch, S. (2002). Lighting up the internet: The Brown and Williamson collection. In R. J. Cox (Ed.), *Archives and the public good: Accountability and records in modern society* (pp. 135–162). Quorum Books.

Elings, M. (2016, May 24). *Using NLP to support dynamic arrangement, description, and discovery of born digital collections: The ArchExtract experiment.* Retrieved from bloggERS, Society of American Archivists, Electronic Records Section: https://saaers.wordpress.com/2016/05/24/using-nlp-to-support-dynamic-arrangement-description-and-discovery-of-born-digital-collections-the-archextract-experiment/

Glantz, S. A., Slade, J., Bero, L. A., Hanauer, P., & Barnes, D. E. (1998). *The cigarette papers*. University of California Press.

Gologorskaya, O., Kaplan, L., Cahn, N., Mills, H., & Panahiazar, M. (2019, September 15). *Word usage evolution.* Retrieved from Summer of NLP 2019: https://wiki.library.ucsf.edu/display/NLPBiomed/Summer+of+NLP+2019

Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal*, *30*(2), 155–174. doi:10.1108/RMJ-09-2019-0055

Ilieva, P. (2017, January 20). *UCSF Archives and Special Collections receives NHPRC grant.* Retrieved from UCSF Library: https://www.library.ucsf.edu/news/nhprc-awarded-a-grant-to-ucsf-archives-and-special-collections/

Kronk, C. A., & Dexheimer, J. W. (2020). Development of the gender, sex, and sexual orientation ontology: Evaluation and workflow. *Journal of the American Medical Informatics Association: JAMIA*, *27*(1), 1110–1115. doi:10.1093/jamia/ocaa061 PMID:32548638

Lee, C. (2018). Computer-assisted appraisal and selection of archival materials. *2018 IIIE International Conference on Big Data (Big Data)*, 2721-2724. 10.1109/BigData.2018.8622267

Maekdo, K. (2019). *About Black Lesbian Archives.* Retrieved from Black Lesbian Archives: https://blacklesbianarchives.wixsite.com/info/about

Mani, N. S., Cawley, M., Henley, A., Triumph, T., & Williams, J. M. (2021). Creating a data science framework: A model for academic research libraries. *Journal of Library Administration*, *61*(3), 281–300. doi:10.1080/01930826.2021.1883366

Mathew, M., Karatzas, D., & Jawahar, C. V. (2021, January 5). *DocVQA: A dataset for VQA on document images.* https://arxiv.org/pdf/2007.00398.pdf

Padilla, T. (2019). *Responsible operations: Data science, machine learning, and AI in libraries*. OCLC Research.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019, May 22). *Final report - always already computational: Collections as data (version 1).* https://zenodo.org/record/3152935

Proctor, R., & Risi, S. (n.d.). *Documentation*. Retrieved from Tobacco Analytics: http://www.tobacco-analytics.org/documentation

Risi, S. (n.d.). *Tobacco networks: The industry's correspondence networks at a glance*. Retrieved from Tobacco Networks: https://tobacconetworks.dhmit.xyz/

Risi, S., & Proctor, R. N. (2020). Big tobacco focuses on the facts to hide the truth: An algorithmic exploration of courtroom tropes and taboos. *Tobacco Control*, (29), e41–e49. PMID:31519796

Schmidt, H., Butter, K., & Rider, C. (2002). Building digital tobacco industry document libraries at the University of California, San Francisco Library/Center for Knowledge Management. *D-Lib Magazine: the Magazine of the Digital Library Forum*, *8*(9). Advance online publication. doi:10.1045eptember2002-schmidt

Shallcross, M., & Prom, C. (Eds.). (2016). *Appraisal and acquisition strategies*. Society of American Archivists.

Thombley, R., Renslo, J., Tonner, C., & Gopez, A. (2019, September 15). *Geographic hotspots for activism during the HIV/AIDS crisis.* Retrieved from Summer of NLP 2019: https://wiki.library.ucsf.edu/display/NLPBiomed/Summer+of+NLP+2019

UCSF Archives & Special Collections. (2021). *AIDS history project.* Retrieved from UCSF Library: https://www.library.ucsf.edu/archives/aids/

UCSF Industry Documents Library. (2021, November 24). *Tobacco litigation documents*. Retrieved from UCSF Industry Documents Library: https://www.industrydocuments.ucsf.edu/tobacco/research-tools/litigation-documents/

United States v. Philip Morris USA, Inc., 449 F. Supp. 2d 1 (D.D.C. 2006).

University of California Libraries. (2017, August 29). *Systemwide annual plan and priorities.* Retrieved from University of California: https://libraries.universityofcalifornia.edu/groups/files/about/docs/FY17-18_AnnualPlanAndPriorities_Final.pdf

University of Georgia. (2005). *UGA tobacco document corpus and toolkit*. Retrieved from NIH-NCI Tobacco-Documents Project at The University of Georgia: http://tobaccodocs.galib.uga.edu/

ADDITIONAL READING

Anderson, S. J., McCandless, P. M., Klausner, K., Taketa, R., & Yerger, V. B. (2011). Tobacco documents research methodology. *Tobacco Control*, (20), ii8–ii11. PMID:21504933

Bero, L. (2003). Implications of the tobacco industry documents for public health and policy. *Annual Review of Public Health*, *24*(1), 267–288. doi:10.1146/annurev.publhealth.24.100901.140813 PMID:12415145

Blummer, B., & Kenton, J. M. (2018). Big data and libraries: Identifying themes in the literature. *Internet Reference Services Quarterly*, *23*(1-2), 1–2, 15–40. doi:10.1080/10875301.2018.1524337

Center for International and Environmental Law. (2017). *Smoke and fumes: The legal and evidentiary basis for holding big oil accountable for the climate crisis*. Retrieved from https://www.ciel.org/wp-content/uploads/2019/01/Smoke-Fumes.pdf

Chicago Tribune. (2012). *Playing with fire*. Retrieved from Chicago Tribune: http://media.apps.chicagotribune.com/flames/index.html

Federer, L., Clarke, S. C., & Zaringhalam, M. (2020, January 16). Developing the librarian workforce for data science and open science. doi:10.31219/osf.io/uycaxosf.io/uycax

LeClere, E. (2018). Breaking rules for good? How archivists manage privacy in large-scale digitisation projects. *Archives and Manuscripts*, *46*(3), 289–308. doi:10.1080/01576895.2018.1547653

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing.

Shanks, J. D., Mannheimer, S., & Clark, J. A. (2019, November 20). *Radical collaboration: Making the computational turn in special collections and archives*. ACRL Digital Scholarship Section Research-DataQ Editorial. Retrieved from https://researchdataq.org/editorials/radical-collaboration-making-the-computational-turn-in-special-collections-and-archives/

KEY TERMS AND DEFINITIONS

Application Programming Interface (API): A type of software interface which enables the exchange of data between systems. APIs generally follow a set of standards documented in an API specification, which provides a user with instructions of how to build or use the API connection. APIs are frequently used with web applications to share large amounts of data for public use.

Discovery Documents: In the United States, parties in a lawsuit may request documents and other information from the opposing party which may be relevant to the case. This pre-trial stage of litigation is called discovery. Documents produced in discovery may be used as evidence if the case goes to a trial.

Library Carpentry: A non-profit community organization which helps to build software and data skills in a library and information science context and to empower people to use software and data in their work.

Master Settlement Agreement (MSA): A 1998 legal agreement entered into by the four major tobacco companies and 46 U.S. States, which resolved the States' lawsuits against the tobacco industry for recovery of tobacco-related health-care costs. The MSA required the companies to end certain marketing practices, to pay over $206 billion to the States, and to make their internal documents available to the public.

Natural Language Processing (NLP): A form of artificial intelligence specifically focused on designing and using computer software, systems, and code that allow a computer to process and "understand" text and spoken words in the same way that humans do.

Optical Character Recognition (OCR): The electronic conversion of an image of text into machine-readable text. OCR text is generated from scanned documents using software. The text of a document can then be used for text mining or other computational analysis, or to provide full-text search capability.

Solr Server: A full-text open-source search platform developed by the Apache Solr Foundation. Its functions include highlighting search terms in results, faceted searching, real-time indexing, database integration, and handling of Word and PDF documents.

Topic Modeling: A text mining tool used to analyze words in a document (or collection of documents) to discover frequently used terms and to group them into clusters. These clusters can provide insight into the topics of a document or collection, allowing a user to better understand their content without needing to read through thousands or millions of pages.