



Using AI and ML to optimize information discovery in under-utilized, Holocaust-related records

Kirsten Strigel Carter¹ · Abby Gondek² · William Underwood³ · Teddy Randby⁴ · Richard Marciano⁵

Received: 30 April 2021 / Accepted: 23 November 2021

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2022

Abstract

Digital cultural assets are often thought to exist in separate spheres based on their two principal points of origin: digitized and born digital. Increasingly, advances in digital curation are blurring this dichotomy, by introducing so-called “collections as data,” which regardless of their origination make cultural assets more amenable to the application of new computational tools and methodologies. This paper brings together archivists, scholars, and technologists to demonstrate computational treatments of digital cultural assets using Artificial Intelligence (AI) and Machine Learning (ML) techniques that can help unlock hard-to-reach archival content. It describes an extended, iterative study applied to digitized and datafied WWII-era records housed at the FDR Presidential Library, rich content that is regrettably under-utilized by scholars examining American responses to the Holocaust. Authors detail the benefits of interdisciplinary collaboration for evaluating user needs, identifying and applying tools and methodologies (including ML through object detection and AI through Named Entity Recognition or NER), and reaching the real-world outcome of public access to augmented data. They also discuss issues of digital representation, relational context, and interface design to enable new modes of public and scholarly access. While based on a case study, we believe that this work is a substantial contribution to revealing the strengths and weaknesses of using AI/ML systems in cultural organizations. We give particular care to lessons learned, and generalize the approach taken across broad classes of collections with a focus on responsive iterations, reproducibility, and the relevance of data and its structures to users.

Keywords FDR Presidential Library · Morgenthau Holocaust Collections Project · Machine learning · Digital curation · User interface development

1 Purpose

This paper describes the collaborative, interdisciplinary research and development undertaken to apply computational thinking and emerging digital methodologies to significant Holocaust-era archival material at the Roosevelt Library. The extensive Diaries of U.S. Treasury Secretary,

Henry Morgenthau, Jr.—864 bound volumes in total, representing 12 years of daily business records—provide a focused, yet complex, textual corpus for these experimental approaches. In this paper, we illustrate our efforts, including trial and error, to extract and augment data drawn from the Morgenthau Diaries’ original custom indexing system first developed and used by the Secretary’s own office, combined

✉ Kirsten Strigel Carter
kirsten.carter@nara.gov

Abby Gondek
abbygondek@outlook.com

William Underwood
underwod@umd.edu

Teddy Randby
tedrandby@gmail.com

Richard Marciano
marciano@umd.edu

¹ Franklin D. Roosevelt Presidential Library and Museum, Hyde Park, NY, USA

² Roosevelt Institute and Franklin D. Roosevelt Presidential Library and Museum, Hyde Park, NY, USA

³ College of Information Studies, University of Maryland, College Park, MD, USA

⁴ Advanced Information Collaboratory (AIC), University of Maryland, Durham, NC, USA

⁵ Advanced Information Collaboratory (AIC), University of Maryland, College Park, MD, USA

with more recent digitized image data and descriptive structures imposed by archivists. Our datafication and augmentation efforts will culminate in improved access and usability outcomes for the public and data scholars alike. We assert that such improvements engender deeper human understanding of the collection, will inspire and sustain new and more complex historical inquiries, and unlock potential for new modes of interpretive analysis.

While the use of Artificial Intelligence (AI) and Machine Learning (ML) in historical archives is in its exploratory stage (Colavizza et al. 2021; Cordell 2020), it represents an emerging trend as seen by its investigation by a number of cultural organizations, including: Yad Vashem, the World Holocaust Remembrance Center, The Smithsonian Institution Data Science Lab in collaboration with the United States Holocaust Memorial Museum (USHMM), the European Holocaust Research Infrastructure (EHRI) project, the Library of Congress “Newspaper Navigator” Dataset, The National Archives (UK), and now many others. On April 15, 2021, the US National Archives (NARA) announced the release of a number of open datasets for AI/ML processing experimentation (8 datasets with object-level metadata from 3000 to 100,000 objects, 4 datasets without object-level metadata from 7000 to 25,000 objects, 2 born-digital datasets with less than 100 objects, and finally 2 datasets with photographic records of people with 500 and 8000 objects, and one nineteenth century portrait collection of notable people with less than 100 objects).

To respond to this challenge, in Feb. 2020, Marciano, Underwood and colleagues launched the Advanced Information Collaboratory (AIC) (<https://ai-collaboratory.net>) with partners from leading academic and cultural institutions, with an emphasis on exploring the opportunities and challenges of “disruptive technologies” for archives and records management [Artificial Intelligence (AI), Machine Learning (ML), Computational Archival Science (CAS), etc.], and leveraging the latest technologies to unlock the hidden information in massive stores of records. More recently, the AIC launched a targeted AI / ML/CAS initiative called the Future of Archives and Records Management (FARM).

This paper represents the first *FARM Initiative* case study exploring the use of Machine Learning strategies to apply predictive modeling in the extraction of valuable hidden archival materials, using the *Morgenthau Holocaust Collections Project (MHCP)*, a privately funded project of the Roosevelt Institute (RI), the nonprofit partner to the Franklin D. Roosevelt Presidential Library & Museum, the first of 15 Presidential Libraries operated by the National Archives and Records Administration. The Morgenthau Holocaust Collections Project at the FDR Presidential Library and Museum is a digital history and path-finding initiative to raise awareness of the Library’s unique but under-explored resources for Holocaust Studies.

We first introduce the Morgenthau Diaries collection and its current public interfaces that center original index and finding tools developed by the record creators. We then discuss the use of the National Archives Catalog API and the FRANKLIN finding aids database to harvest digital content (files and metadata) to drive the creation of a new, global master index for all 864 volumes in the collection using a 2-Phase Supervised Machine Learning algorithm based on object detection methodologies: “Object Detection enables developers to train custom machine learning models that are capable of detecting individual objects in a given image along with its bounding box and label.”¹ Beyond ML, we establish the value and suitability of implementing “*document object-schema databases*” for historical subject index content, augmenting this content through the use of AI NLP/NER methodologies (Qi et al. 2020), the need for bringing humans back in the loop for verification and validation of ML automated content, new or next approaches for querying the validated database, and the need to further consider issues surrounding representational equity. This thread leads to a discussion on interface design to enable new modes of public and scholarly access, replicability, and scalability, and finally, we conclude with a discussion of future directions for iterations of applied AI and ML treatments, and the significance of lessons learned in this effort.

2 Introduction to the Morgenthau Diaries and the Morgenthau Holocaust Collections Project

The Morgenthau Holocaust Collections Project (MHCP) of the FDR Presidential Library and its nonprofit partner the Roosevelt Institute is a grant-funded scholarly initiative to explore and develop information pathways for better use of the FDR Library’s Holocaust-related archival materials. The project is launched in 2017 (FDRL 2021), its first content focus being the Diaries and Papers of Henry Morgenthau Jr. (HMJr) who was an influential Cabinet member during both the Roosevelt and Truman administrations, 1934–1945. Henry Morgenthau Jr. and Franklin D. Roosevelt were good friends and Hudson Valley neighbors, and HMJr is considered by most historians as a key figure in the American government’s response to the Holocaust (FDRL2018a; b).

A central goal of the MHCP is to improve and promote new uses of archival materials in Holocaust research projects of all kinds. Prior to the experimental treatments

¹ Google AutoML Vision Object Detection. See: <https://cloud.google.com/vision/automl/object-detection/docs>

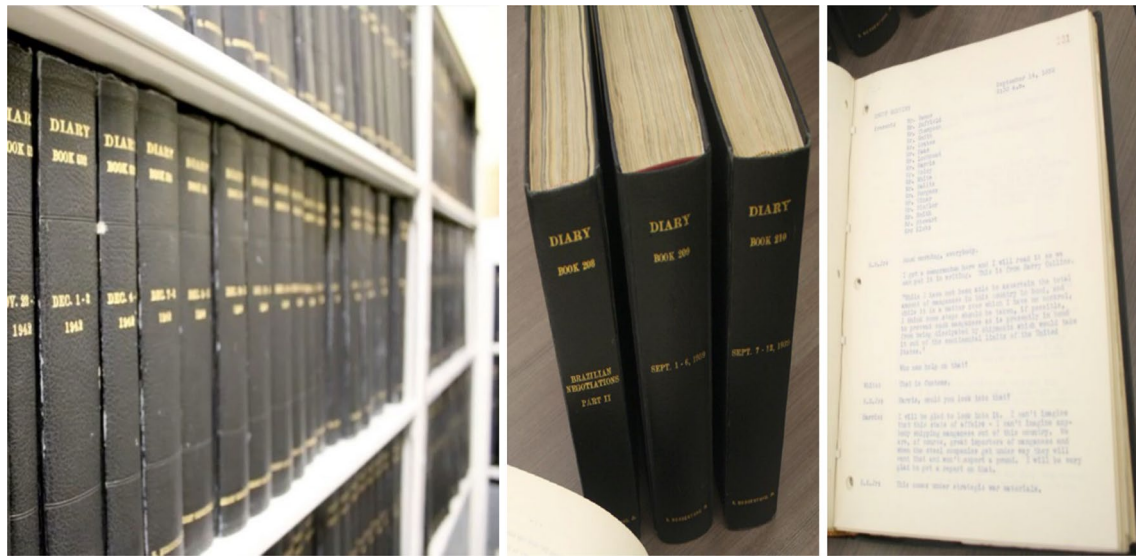


Fig. 1 Volumes 208, 209, and 210

described in this paper, the MHCP supported three major thematic explorations by the current Morgenthau Scholar-in-Residence, Abby Gondek: (1) the impact of lesser known figures, especially women like Henrietta Klotz (Henry Morgenthau Jr.'s secretary), on the actions taken by the Treasury Department and the War Refugee Board (WRB) to bring relief and rescue to Jewish refugees (Gondek 2020a); (2) quantifying patterns in letters from the public in support of “free ports” in the US in 1944 (with a special focus on letter writing campaigns implemented by Jewish women’s organizations) and the role of the WRB in the establishment of the Oswego “Emergency Refugee Shelter” (Gondek 2021); (3) the conflicts between Treasury and State (and Jewish organizations) regarding rescue of Jewish children from France in 1943–1944 (Gondek 2020b). These prior inquiries helped set a historical/interpretive frame of reference for further queries in our newly computational environment.

2.1 The Morgenthau Diaries

During Henry Morgenthau, Jr.’s nearly 12 years as FDR’s Secretary of the Treasury, he compiled more than 860 diary volumes. These are not typical diaries; rather, they are Morgenthau’s daily record of his official activities, including transcripts of his meetings and telephone conversations as well as copies and originals of the most important correspondence and memoranda that passed over his desk. Because of Morgenthau’s long tenure in government and his close personal relationship with FDR, the Diaries also document Morgenthau’s involvement and interest in New Deal fiscal and monetary policy during the Great Depression, wartime economic mobilization and aid to the Allies,

post-war planning and the so-called “Morgenthau Plan” for Germany, the plight of European Jews, creation of the War Refugee Board, planning for the Bretton Woods and United Nations conferences, as well as other social, political, economic, and diplomatic issues. Combined, these Diaries and companion subject index cards total some 285,000 pages (FDRL 2018b).

Henrietta Klotz served as HMJr.’s primary personal secretary for 37 years, managing his schedule and effectively running his office (NYT Archive 1988; Morgenthau Jr. 1945). Klotz supervised the curation and indexing of the Henry Morgenthau, Jr., Diaries (Gondek 2020a; FDRL 2018a) which were bound in chronological order. Each volume typically covered 1–3 days and included key correspondence, memos, and meeting and telephone transcripts. Treasury Librarian Isabella Diamond supervised the binding process, including systematic subject indexing first for Morgenthau’s press conferences in 1936, and later for the Diaries in 1939 (Diamond 1941, 1943; Gondek 2020a; McReynolds 1939; Morgenthau Jr. 1936; Morgenthau III 1991). Under Diamond’s direction, microfilming began in August 1941, and it proved a time-consuming task. At 500 pages per book, microfilming two volumes took an entire day to complete (McReynolds 1939; Diamond 1941, 1943). Henry Morgenthau Jr. complimented her work: “I am struck with the care and precision with which the indexing is done... It will always be easy for me to find my way around in the volumes because of this clear and thorough indexing” (Morgenthau Jr. 1936). As of 17 December 1943, 645 volumes (books) through June 1943 were indexed and 600 volumes through December 1942 were bound (Fig. 1).

Fig. 2 TOC pages for Volume 696 (January 22–26, 1944)

In 1943, Morgenthau Jr. became involved in the debate over Jewish refugees; he was foundational in the development of the War Refugee Board, which led to the rescue of at least 200,000 Jews from Nazi-occupied countries (FDR Library 2018b). Interestingly, he was also a secular, assimilated Jew who had never even attended a Passover seder. Influenced by his father's emphasis on being "American" and not Jewish, HMJr was not a Zionist, and avoided "Jewish matters." Henry Morgenthau Sr. was a U.S. Ambassador to the Ottoman Empire in Turkey during the Armenian genocide. He dedicated himself to raising awareness and funds to stop the "race extermination" of the Armenian people. His father's legacy inspired Henry Morgenthau Jr.'s decision to take action on the part of Jews persecuted by Nazism. Henrietta Klotz was extremely influential in convincing Morgenthau to take a more active stance in regard to rescuing Jews (Adalian 2019; Beschloss 2002; Erbelding 2015; Klotz 1986, Morgenthau Jr. 1945; Morgenthau III 1991). After HMJr.'s forced resignation during the Truman Administration in 1945, he became very involved in Jewish causes through Henrietta Klotz's social, philanthropic, and service networks; he became the chairman of the United Jewish Appeal and financially advised the state of Israel (Gondek 2020a; Erbelding 2015; Morgenthau Family Papers n.d.; NYT Archive 1988; Penkower 2016).

The Morgenthau Diaries represent a unique corpus of high intrinsic value, especially with regard to analyzing daily happenings of the U.S. government during the Roosevelt Administration. Scholars of Holocaust Studies have long known of the collection and used its most widely known components to form arguments interpreting American responses, but given the sheer depth and detail of the information available within its volumes, without digital assistance in both finding and usability, the source remains underexplored. Viewing this material through a lens of collections as data (Padilla et al. 2019), and with funding through the

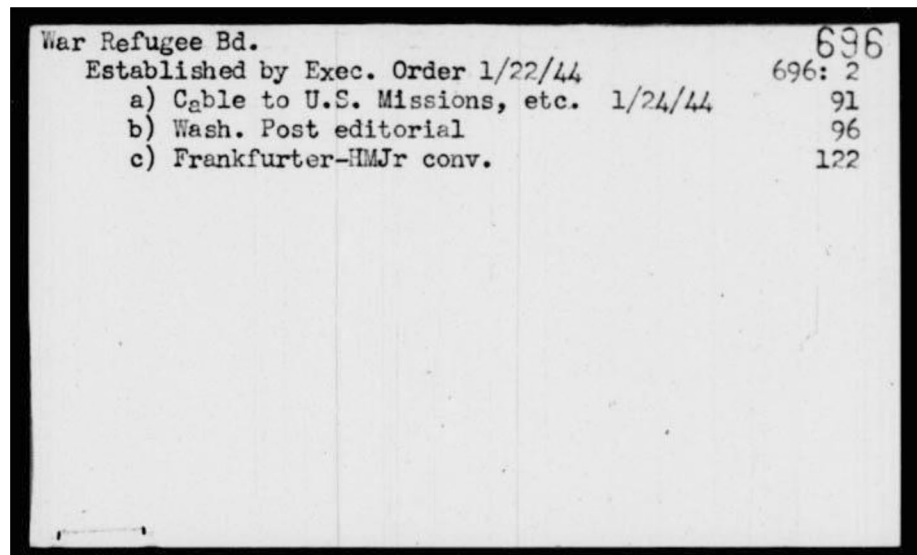
Morgenthau Holocaust Collections Projects (MHCP) in partnership with the University of Maryland Advanced Information Collaboratory (AIC)/FARM Project, the FDR Library seeks to enhance access to this key primary source, and to realize its potential for both digital scholarship and greater public understanding of the Holocaust.

2.2 Analyzing the original table of contents (TOC) and index cards

The FDR Library digitized the historical microfilm version of the Morgenthau Diaries in its entirety in 2014. These microfilm scans were processed into compressed multi-page PDF access files, digitally arranged to match the physical collection, and posted online in traditional finding aid and catalog settings. This approach allowed the Library to publish the full digital collection expediently, with staff using simple batch scanning software for capture, Adobe Photoshop Lightroom for image editing and arrangement, and Adobe Acrobat Pro for both PDF creation and OCR. Rather than itemizing and creating document-level or text-level metadata, staff utilized existing collection description (both finding aid and NARA Catalog records down to the file unit level) for initial web publishing. Our research team used this publicly available digital collection of scanned microfilm for its further computational treatments.

As described in Sect. 2, the physical Diary Volumes are arranged chronologically, with documents bound in leather book-style covers. Both the historical microfilm and its online digital surrogates mimic this arrangement. Each Volume begins with a title page showing the Volume number and the Volume's inclusive dates. Title pages are followed by a Volume-specific table of contents (TOC), and each Volume has its own pagination beginning with page one. Figure 2 shows the title page and the 4 pages of the TOC for Volume 696. TOCs hold a great deal of Volume-specific

Fig. 3 First of six index cards referencing Vol. 696



structured subject classification information, and so function more like a standalone subject index than a traditional table of contents. The TOCs reflect the various subject headings, sub-headings, and cross-references assigned by Isabella Diamond according to her custom schema. Volume numbers, page numbers, and dates of component documents are also indicated. Often, these subject and document references direct the reader to a future Volume number with associated pages and dates.

Volume 696 is historically significant, because it contains documents from January 22, 1944 when the War Refugee Board was created through executive order 9417. The Executive Order declared “it is the policy of this Government to take all measures within its power to rescue the victims of enemy oppression who are in imminent danger of death and otherwise to afford such victims all possible relief and assistance consistent with the successful prosecution of the war.” The Board was made up of the Secretaries of State, Treasury, and War and was tasked with providing rescue and relief to these victims and “the establishment of havens of temporary refuge for such victims” (Roosevelt 1944a). The War Refugee Board was a result of 2 years of advocacy on the part of World Jewish Congress representatives like Gerhart Riegner, Emergency Committee to Save the Jewish People of Europe leader Peter Bergson, congress members Guy Gillette and Will Rogers Jr., and Morgenthau Jr.’s Treasury staff (including Josiah DuBois, Randolph Paul, and John Pehle). These individuals and organizations were responding to the inaction and prevention of action in the State Department regarding relief and rescue of refugees (Breitman and Lichtman 2013; Erbelding 2018; Gondek 2020b; Morgenthau Jr. et al. 1944; Paul 1944).

“War Refugee Board” first appears in Diamond’s classification in Volume 696. Previous headings included “Refugees” or “Refugees (Jewish).”

Using the same custom classification scheme, Diamond created an index card file to provide document-level access across Volumes by subject. The index cards are alphabetically arranged, with each subject entry citing a Diary Volume number, the relevant page numbers for component documents, document dates, and occasional cross-references to other Volumes indicated by a “see also” note. For example, Fig. 3 shows 696:2, which stands for *Volume 696* at page 2.

Figure 4 shows the entire “War Refugee Board” TOC entry for Volume 696, which spans the last two pages of the TOC for *Volume 696*, and how its contents relate directly to individual index cards. Note that some of the index card entities appear in handwritten script, including in-line edits. These amendments are contemporary to the records, and were not added by FDRL archivists.

2.3 Current web interfaces

We used both NARA’s Catalog [API](#) and also the current Morgenthau Diaries finding aid hosted in the FDR Library’s FRANKLIN finding aids web interface to download all 864 Volumes and their associated metadata:

- NARA’s API (Application Programming Interface) allowed us to write a simple Python script (using Jupyter Notebooks), where starting at the Collection level (Identifier: “FDR-MORGEN”), we walked down the hierarchical archival description tree to the Diaries of Henry Morgenthau, Jr. Series (Series Identifier:

- W -		Book	Page
War Refugee Board			
Established by Executive Order January 22, 1944.....		2	
a) Cable to United States Missions, etc. - 1/24/44.....		91	
b) Washington Post editorial.....		96	
c) Frankfurter-HMJr conversation.....		122	
Jeffers, William, suggested by Rosenman as Executive Director of War Refugee Board			
Frankfurter-HMJr conversation - 1/26/44.....		123	
9:30 group conference - 1/26/44.....		127	
a) "Relief to Jews in Roumania, Slovakia, Croatia, and Theresienstadt through International Committee of Red Cross" - memorandum.....		150	
b) "Evacuation of Jews from Poland to Hungary" - memorandum.....		156	
c) "Evacuation of Jews from Transnistria to Roumania" - memorandum.....		158	
d) "Evacuation of Abandoned Children from France to Switzerland" - memorandum.....		159	
e) "Evacuation of Yugoslav Refugees on the Island Sab to Italy" - memorandum.....		161	
f) Cable explaining War Refugee Board sent to United States Missions at Lisbon, Madrid, Stockholm, Bern, and Ankara.....		164	
(See also Book 692, page 178)			
1) War Department cooperation with representatives abroad offered:			
See Book 692, pages 168-181			
2) British War Office may be asked to Board's first meeting reported by HMJR to White, Pehle, Luford, and DuBois - 1/26/44.....	696	168	
a) State Department wants board entirely separate from State			
b) Alternates: Stettinius for Hull; Gaston for HMJR; Bundy or McCloy for Stimson			
c) Pehle: HMJR advocates as Executive Director....		168	
d) Hull's agenda		178	
Conference of Treasury group - 1/26/44.....		183	
a) FDR-HMJr conversation reported			
b) Bruere, Murray, Bullitt, Willkie, Hamilton Fish Armstrong discussed			
c) Stassen (of Minnesota) suggested		202, 221	
1) Dodge, Cleveland H., suggested.....		221	
2) Hoffman, Paul			
Johnston, Eric			
(See also Book 692, page 20)			
d) Ernst (Morris)-HMJR conversation.....		211	
1) Stassen (of Minnesota) suggested			
(See also Book 692, page 39)			
e) Wise (Stephen S.)-HMJR conversation.....		218	
1) McDonald (James G.) suggested			
(See also Book 692, page 38)			
* Executive Director suggestions			
Stimson letter - 1/24/44.....		229-A	
Tracy letter - " (See also Book 692, page 38).....		229-B	

Fig. 4 Correspondence between the TOC subject heading "War Refugee Board" and index cards for Volume 696

589213), and then to its 879 chronologically arranged File Units, finally downloading some 1167 associated digital objects.

- FRANKLIN's Morgenthau Diaries interface allowed us to play with a different technique, crawling the finding aid page, downloading all the objects referenced within, and re-creating a content tree in the cloud (we store it in Box). Rather than extracting the metadata from NARA's API, which was an option, we instead used a hybrid approach and extracted the metadata from the FRANKLIN web crawl.

The FRANKLIN finding aids database presents the digitized Morgenthau Diaries collection description in archival context. It features hyperlinks to PDF versions of scanned microfilm associated with Volume titles in a simple list. This user experience is designed to mimic an in-person research room experience whereby a user selects a particular file unit and proceeds to use it exclusively, or at least one resource at a time.

Using the FRANKLIN interface, a typical researcher query would go as follows. If interested in the "War Refugee Board:"

- Navigate to the finding aid for the Morgenthau Diaries and select "Series 3: Morgenthau Diaries Index." There are 166 Index Cards referencing this topic, ranging from Volume 688 (May 7-Dec 8, 1943) to Volume 837 (Apr. 13-16, 1945). The 5th Index Card with the heading, "War Refugee Bd." notes "Establishment by Exec. Order on 1/24/44" and points the reader to Volume 696/Page 2. See Fig. 3 for Index Card display.
- Using this lookup information, one would then navigate to the list of Volumes presented in the Morgenthau Diaries finding aid as "Series 1: Morgenthau Diaries". The Volumes are listed online (FDRL 2018a) starting from 00 to 864, as shown in Fig. 5. One would then select the "View Online" option for Volume 696, which would display the entire Volume as a multi-page PDF (all 302 pages) (Fig. 6).

Volume 00, Farm Credit Diary, April 27–November 16, 1933 [View Online](#)
 Volume 0, London International Monetary and Economic Conference, June, 1933 [View Online](#)
 Volume 1, October 1933–June 30, 1934 [View Online](#)
 Volume 2, July 1–December 31, 1934 [View Online \(Part 1\)](#); [View Online \(Part 2\)](#)
 Volume 3, January 1–February 28, 1935 [View Online \(Part 1\)](#); [View Online \(Part 2\)](#)
 Volume 4, March 1–April 22, 1935 [View Online \(Part 1\)](#); [View Online \(Part 2\)](#); [View Online \(Part 3\)](#)
 Volume 5, April 23–May 30, 1935 [View Online \(Part 1\)](#); [View Online \(Part 2\)](#)
 Volume 6, June 1–June 15, 1935 [View Online](#)
 Volume 7, June 16–June 30, 1935 [View Online](#)
 Volume 8, July 1935 [View Online](#)
 Volume 9, August 1935 [View Online](#)
 Volume 10, September–October 31, 1935 [View Online](#)

.....

Volume 696, January 22–January 26, 1944 [View Online](#)

.....

Volume 854, June 11–June 13, 1945 [View Online](#)
 Volume 855, June 14–June 18, 1945 [View Online](#)
 Volume 856, June 19–June 21, 1945 [View Online](#)
 Volume 857, June 22, 1945 [View Online \(Part 1\)](#); [View Online \(Part 2\)](#)
 Volume 858, June 23–June 26, 1945 [View Online](#)
 Volume 859, June 27, June 29, 1945 [View Online](#)
 Volume 860, June 30, 1945 [View Online](#)
 Volume 861, July 1–July 3, 1945 [View Online](#)
 Volume 862, July 4–July 9, 1945 [View Online](#)
 Volume 863, July 10–July 14, 1945 [View Online](#)
 Volume 864, July 14–July 21, 1945 [View Online](#)

Fig. 5 Online box and folder listing for series 1: Morgenthau Diaries

The original custom indexing schema established broad subject-based control across the diary volumes which were otherwise bound to strict chronological arrangement. It served as the first, and most heavily used, finding system for the Diaries, while those records were still in active use by the Treasury Secretary’s office. The existing modern day finding aid and catalog interfaces convey that original context faithfully. While those systems were and remain extremely valuable for understanding the provenance and function of the original corpus, new computational approaches to content discovery are now needed to support fuller, deeper access to the collection. Datafying the content and augmenting its indexing can support new and further computational uses of the Diaries for any research goal.

3 Re-creating a master index through machine learning

We developed a 2-phase Supervised Machine Learning (ML) algorithm for automatically re-creating a global index across all 864 Volumes, using the data extracted from the hybrid NARA Catalog API/FRANKLIN finding aid extraction

process described in Sect. 2.2. The training we conducted to produce these results literally consisted of drawing boxes on a subset of the TOCs and labeling those with a “box” tag for Phase 1, and as 4 other types of tags for Phase 2 (details in Sect. 3.1). The tools used include *Google AutoML Vision “Object Detection”* and *nanonets.com*. The approach is detailed in a previous paper (Randby and Marciano 2020). This 2-phase ML pipeline processes a total of 3579 TOC JPG pages, discovers a total of 22,667 Subject Index boxes (an average of 6 to 7 boxes per TOC page), and produces a final database of indexes. We extracted 3 indexes: *Subject-Headers* (the titles you see on individual Index Cards), *Volume-Page* references in the TOCs, and *Dates* in the TOCs.

Typical uses of *Object Detection models* are for drawing boxes around apples, flowers, or cars in images, and *nanonets.com* for capturing content from images of invoices, receipts, and driver’s licenses. We believe our demonstration of new uses of these kinds of industry tools toward valuable historical and cultural analysis to be both innovative and original.

The microfilm developed under Isabella Diamond’s supervision in the 1940s was used to create the scanned images of the Diaries. These images in pdf form offer poor readability and include many OCR errors. The methods

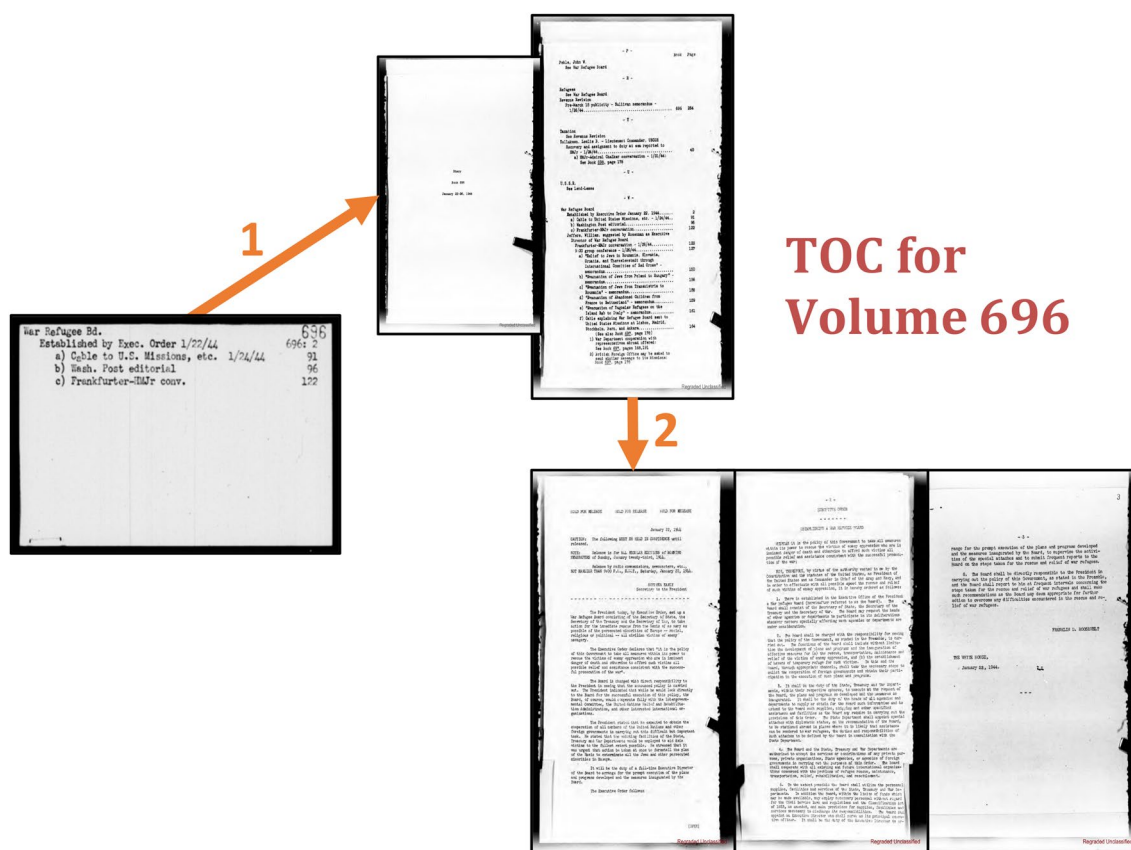


Fig. 6 Illustration of the finding aid navigation process

described in this paper are an attempt to obtain information from the TOCs without correcting the OCR errors for 285,000 pages, estimated to take 35 years of human labor.

3.1 Phase 1 of the ML extraction: finding subject header boxes

Phase 1 involved training to recognize subject index headers within TOC pages. In Fig. 7, we show the subject index box labels discovered by the Phase 1 ML model for *Volume 696*. There are 31 subject index boxes with the following headers:

Argentina / Armstrong / Hamilton Fish / Baerwald, Paul / Bolivia / Bridgeport / Connecticut, Bruere, Henry / Bullitt, William C. / Business Conditions / China / Connecticut, Bridgeport / Dakar / Ernst, Morris / Financing, Government / Foreign Funds Control / France / Latin America / Lend-Lease / Mann, James / War Refugee Board / National Gallery of Art / Pehle, John W. / Refugees / Revenue Revision / Taxation / Tollaksen, Leslie B. - Lieutenant Commander, USCGR / U.S.S.R. / **War Refugee Board** / Widener Art Collection / Willkie, Wendell L. / Wise, Rabbi Stephen S.

3.2 Phase 2 of the ML extraction: object detection

Phase 2 of the ML algorithm further drills down into the boxes identified in Phase 1, by detecting 4 classes of objects: *Book/Page* pairs, *Dates*, *Headers*, and lines of *Content*. Object detection models are typically used to draw boxes around apples or cars in images—but they can be very versatile given enough training data. To train the model, a training data set of labeled images was created. The images began as the sub-sections created during phase one, and then, each subsection was annotated by drawing boxes around each

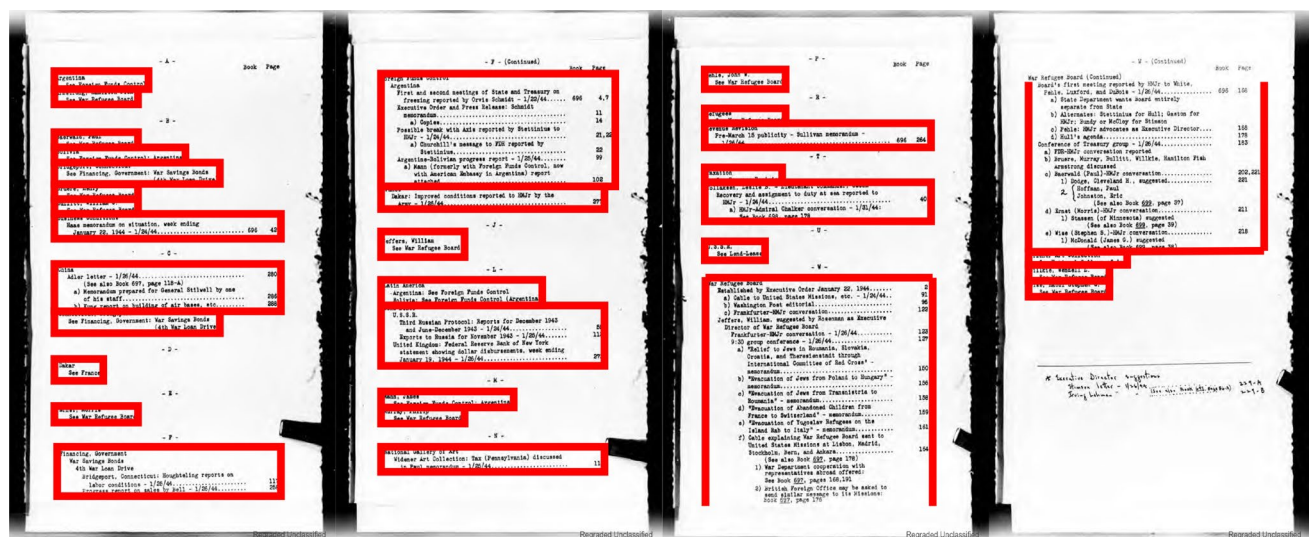


Fig. 7 Subject index boxes discovered by the phase 1 ML model for Volume 696. Results show that for Volume 696, the model identifies 31 Subject Index boxes [An independent application of the ABBYY FineReader Optical Character Recognition (OCR) tool on the

object type. For example, headers in these sub-sections were annotated with a box and labeled “Header.” This process was repeated for each of the four labels on a training set of several hundred sub-sections. Once trained, the model can recognize labels and the area on the page where they occur, as shown in Fig. 8. These labels are then sorted by position from top-left to bottom-right, and fed one by one into a parsing state machine. The output of this process is a new object which contains an OCR text *Header*, some OCR text *Content*, a list of *Book/Page* pairs, and a list of *Dates*. Notice that, in the index, there is an implied relationship between *Content*, *Book/Page* pairs, and *Dates*: a segment of content is followed by an optional *Date*, and then later a *Page*. Taking advantage of this principle in the parsing machine, we can keep track of which segments of content are associated with which page and which date.

An analysis of the resulting *Book/Page* index shows that the “War Refugee Board” header content references volume content located not only in Volume 696 but also in Volumes 697 and 699, demonstrating that this is not a traditional TOC.

The most relevant document to current research within the Morgenthau Holocaust Collections Project cited within this TOC is “Evacuation of Abandoned Children from France to Switzerland” on page 159. Gondek used documents related to this topic in an online exhibit, “Jewish refugee children and the Establishment of the War Refugee

Board, 1943–1944.” The issue of rescuing Jewish children from France played a central role in the establishment of the War Refugee Board, though it is not typically discussed as being one of the causal factors (Gondek 2020b).

4 Beyond ML: document DBs, AI, crowdsourcing, and APIs

Standard measures of how well ML models perform often include precision and recall. For Phase 1 of the ML extraction, our model evaluation showed a Precision of 96.5% and a Recall of 80.1%. *Precision* tells us, from all the test examples that were assigned the “box” label, how many actually were supposed to be categorized with that label. *Recall* tells us, from all the test examples that should have had the “box” label assigned, how many were actually assigned the label.

While precision and recall give an indication of how well a model is capturing information, and how much is being left out, this is not necessarily useful for extracting important cultural data.

Next, we discuss crucial post-processing infrastructure, often overlooked, but indispensable in leveraging predicted model output. We first demonstrate the value of assembling ML output into document object-schema databases, to further verify and validate machine-generated results and augment the ML output with AI-generated content.

War Refugee Board	2
Established by Executive Order January 22, 1944	91
a) Cable to United States Missions, etc. 1/24/44	96
b) Washington Post editorial	122
c) Frankfurter-RMJr conversation	
Jeffers, William, suggested by Rosenman as Executive Director of War Refugee Board	
Frankfurter-RMJr conversation - 1/26/44	123
9:30 group conference 1/26/44	127
a) "Relief to Jews in Roumania, Slovakia, Croatia, and Theresienstadt through International Committee of Red Cross" memorandum	150
b) "Evacuation of Jews from Poland to Hungary" - memorandum	156
c) "Evacuation of Jews from Transnistria to Roumania" - memorandum	158
d) "Evacuation of Abandoned Children from France to Switzerland" - memorandum	159
e) "Evacuation of Yugoslav Refugees on the Island Rab to Italy" - memorandum	161
f) Cable explaining War Refugee Board sent to United States Missions at Lisbon, Madrid, Stockholm, Bern, and Ankara	164
(See also Book 697, page 178)	
1) War Department cooperation with representatives abroad ordered. See Book 697, pages 168, 191	
2) British Foreign Office may be asked to send similar message to its Missions. Book 697, page 178	

Fig. 8 Results of phase 2 of the ML model shown on part of the “War Refugee Board” box with 4 labels

4.1 On the value of document object-schema databases and use of AI

The output of the ML algorithm is shown to be tagged text content (expressed as JSON). JSON stands for JavaScript Object Notation and is a lightweight data-interchange format which is “self-describing” and readable in of itself (a reduced version of XML if you will). JSON data are written as name/value pairs, such as “header”: “War Refugee Board” or “book”: “696”. In JSON, values are one of the following data types: *string*, *number*, *object*,² *array*,³ *Boolean*, or *null*.

² JSON objects are surrounded by curly braces {} and can contain other JSON objects, hence the notion of hierarchy.

³ A flat data structure consisting of elements which can be accessed by their position.

We load the ML-generated JSON content into a cloud NoSQL database called MongoDB. MongoDB is uniquely suited for managing the content we extracted as it is considered to be a document object-schema database, where:

- A document (a confusing term on first encounter!) is the basic unit of data for MongoDB, roughly equivalent to a row in a relational database management system (but much more expressive).
 - For us, documents correspond to the Subject Index box content produced by the 2-phase ML pipeline. Since the ML pipeline generated a total of 22,667 header boxes, from the perspective of MongoDB, we populated the database with 22,667 so-called documents.
- Similarly, a collection can be thought of as the schema-free equivalent of a table (but one can also impose a structure, if so desired).

- A single instance of MongoDB can host multiple independent databases, each of which can have its own collections and permissions.
- Every document has a special key, “*_id*”, that is unique across the document’s collection (Chodorow and Dirolf 2010).

A document schema is a JSON object that allows you to define the shape and content of our documents. We used the following document object-schema:

```
{
  _id:      Object
  book:     String
  header:   String
  content:   String
  dates:    Array
  indexes:   Array
  maxDate:  Object
  minDate:  Object
}
```

A spectacular feature is the ability to extend a document object-schema at any point in time and include new arrays or objects as needed. This is where AI approaches to enrich the ML-generated content come into play. We make use of the well-known Stanford National Language Processing (NLP) software (<https://nlp.stanford.edu/software/>). We chose a recent addition to this software called Stanza (Qi et al 2020). Stanza is a Python NLP toolkit that has a fully pretrained neural model supporting Named Entity Recognition or NER). NER recognizes mentions of particular entity types such as Persons or Organizations in text content.

We used Stanza to extract the following named entities from the **content** field: locations, organizations, and people. Having run this post-ML NER workflow, we then incorporate that new content directly into the MongoDB database by simply extending document object-schema to include these new indexes (highlighted in blue):

We illustrate this schema on the “*War Refugee Board*” entry for Volume 696:

```
{
  "_id": { "$oid": "60350de3cd1f800015d0b8da" },
  "book": "696",
  "header": "War Refugee Board",
  "content": "Established by Executive Order January 22, 1944 See:696:2\n
    a) Cable to United States Missions, etc. - 1/24/44 See:696:91\n
    ...
    Irving Lehman —1/26/44 See:229-B"
  "dates": [
    {
      "day": {"$numberInt": "24"},
      "month": {"$numberInt": "1"},
      "year": {"$numberInt": "44"},
      "stringified": "1/24/44",
      "content": "a) Cable to United States Missions, etc."
    }, ...
  ],
  "indexes": [
    {
      "content": "Established by Executive Order January 22, 1944",
      "book": "696",
      "page": "2"
    }, ...
  ],
  "maxDate": { ... },
  "minDate": { ... },
}
```

-100
-10
-1
Volume 696
+1
+10
+100

-100
-10
-1
Entry 27
+1
+10
+100

Entry
Dates +
Indexes +

Header

War Refugee Board

Content

Hull; Gaston for HMJr: Bundy or McCloy for Stimson

c) Pehle: HMJr advocates as Executive Director See:696:168

d) Hull's agenda See:696:178

Conference of Treasury group - 1/26/44 See:696:183

a) FDR-HMJr conversation reported

b) Bruere, Murray, Bullitt, Willkie, Hamilton Fish Armstrong discussed

c) Baerwald (Paul)-HMJr conversation See:696:202,221

1) Dodge, Cleveland H., suggested See:696:221

2) Hoffman, Paul; Johnston, Eric: See:699:37

d) Ernst (Morris)-HMJr conversation See:696:211

1) Stassen (of Minnesota) suggested. See:699:39

e) Wise (Stephen S.)-HMJr conversation See:696-218

1) McDonald (James G.) suggested. See:699:38

Month

1

Day

24

Year

44

Stringified

1/24/44

Content

a) Cable to United States Missions, etc.

remove

Book

696

Page

2

Content

Established by Executive Order January 22, 1944

remove

Month

1

Day

26

Year

44

Stringified

1/26/44

Content

Jeffers, William, suggested by Rosenman as Executive Director of

remove

Book

696

Page

91

Content

a) Cable to United States Missions, etc.

remove

Book

696

Page

122

Content

c) Frankfurter-HMJr conversation

Update
Reset
Delete

Fig. 9 A crowdsourcing app for verification and validation

{		
_id:	Object	
book:	String	
header:	String	
content:	String	
dates:	Array	
indexes:	Array	
maxDate:	Object	
minDate:	Object	
locations:		Array
organizations:	Array	
people:	Array	
}		

This is the resulting “War Refugee Board” document for Volume 696, originally ML-generated and now augmented with new AI-computed fields (highlighted in blue):

back in the loop. We explore combining automation and human validation through the use of an app.

The *Morgenthau Verifier* is built on top of the MongoDB Morgenthau document database. It not only allows the browsing of entries per Volume, but also allows for editing: (1) the header content, (2) the dates’ values, and (3) the indexes for the book/page content. One can also remove “book” and “dates” sub-entries, but also add new ones. In other words, it implements a full CRUD interface (Create, Read, Update, and Delete).

This app has been used to verify and validate ML the output for 60 of the 864 Volumes so far. We focused on 6 months of Diary content from Jan. 1, 1944 to June 30, 1944 which spans 60 Volumes. This half-year period was chosen as it covers the first 6 months of the War Refugee Board’s operations. Figure 9 is a screen snapshot of the initial crowdsourcing app that was developed.

At this stage, the authors recognized or confirmed many limitations posed by the 1940s index terms and subject assignments, and identified potential benefit in adding future layers of associated controlled vocabulary terms for

```
{
  "_id" : { "$oid":"60350de3cd1f800015d0b8da" },
  "book" : "696",
  "header" : "War Refugee Board",
  "content" : "Established ... Irving Lehman —1/26/44 See:229-B",
  "dates" : [...],
  "indexes" : [...],
  "maxDate" : {...},
  "minDate" : {...},

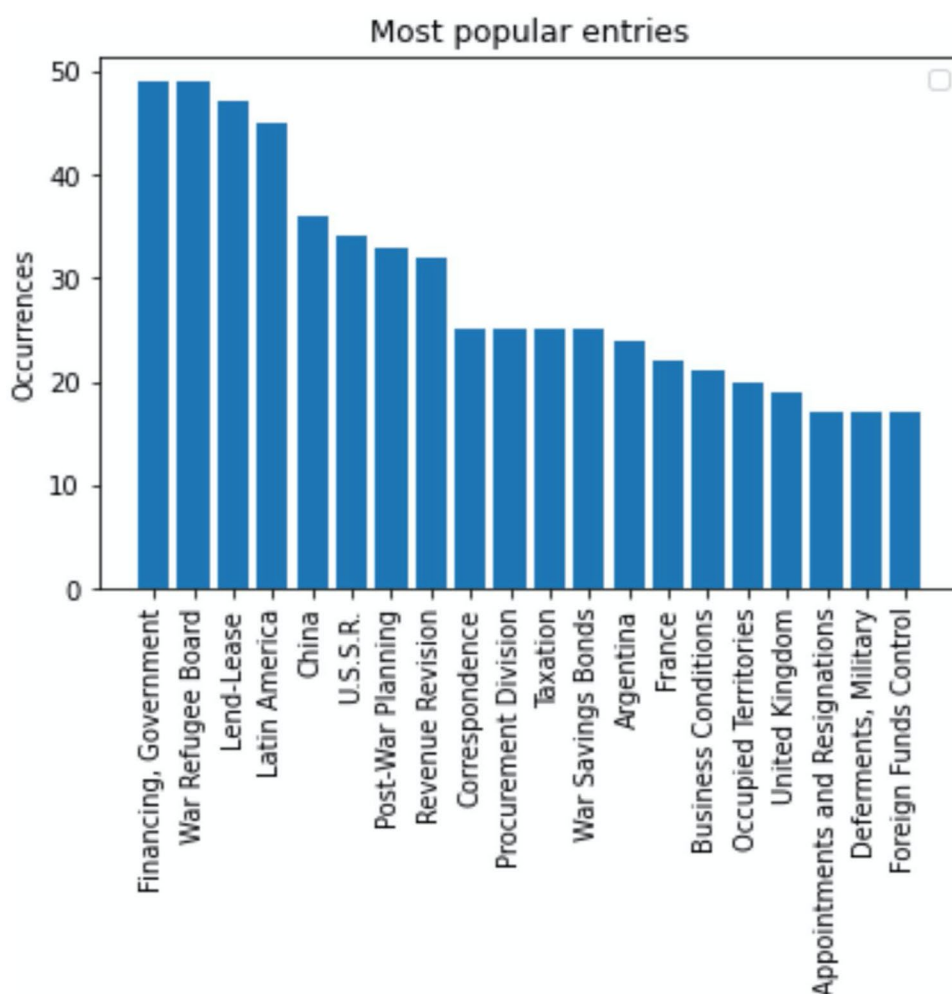
  "locations" : [
    "Roumania", "Slovakia", "Croatia", "Poland", "Hungary", "Lisbon",
    "Madrid", "Stockholm", "Bern", "Ankara", "Minnesota", "Theresienstadt"
  ],
  "organizations" : [
    "Washington Post", "War Refugee Board", "War Department",
    International Committee of Red Cross, "British Foreign Office"
  ],
  "people" : [
    "Jeffers", "William", "Rosenman", "Bundy", "FDR", "Willkie",
    "Hamilton Fish Armstrong", "Paul", "Johnston", "Ernst", "McDonald"
  ]
}
```

4.2 On the need for crowdsourcing for verification and validation

The 2-Phase ML process, however accurate, has the potential of missing key historical content. To mitigate the limitations of automation, it is indispensable to bring the human

the Holocaust and U.S. government during World War II. Some of the terms originally used to describe people, religions, and ethnic groups are offensive, with several biases rooted in a 1940s mentality that also, inevitably, affects the underlying index schema. Applying linked data crosswalks to subject-related content throughout the volumes and the original index would further loosen their restrictive,

Fig. 10 Count of unique headers over the 1st 60 Volumes of 1944 (first 6 months), sorted by highest occurrence



artificial, chronological arrangement, and their initial subject classification, while preserving original context through clear citation and augmented, associative positionality. We also recognized the need for additional, natural language descriptions provided via community/crowdsourced input, important both for improving representational equity in description, and for assistance in identifying, analyzing, and assigning relational connections between document, entity, and even term components. Given that these sources often document, both directly and indirectly, the persecution of Jewish people and other marginalized groups, community perspectives may better “conceiv[e] of records as agents, embodied with the voices of past lives, and capable of facilitating meaning for those who access, activate, and interpret them” (Tai et al. 2019), a goal central to the path-finding mission of the Morgenthau Holocaust Collections Project.

4.3 Toward querying the validated database

As part of our experimentation to explore the suitability of building document databases, we prototyped a collection of

tools for analyzing the Morgenthau Diaries. The initial software is a Python library on top of the MongoDB database and is a first attempt at putting together a package of tools to help researchers with the exploration of the global Diary index (<https://github.com/TeddyRandby/hmjrPyTools>).

We opted to create educational materials in the Jupyter Notebook format. Our choice of this platform is motivated by the fact that “Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document” (Perkel 2018).

This blog-like format allows us to smoothly blend together explanatory text alongside working blocks of example code and the resulting visual output (Marciano et al. 2019). Versions of this library will be demonstrated on our <https://cases.umd.edu> website moving forward.

Example queries include:

Fig. 11 Plotting number of volume entries over time in the first 6 months of 1944

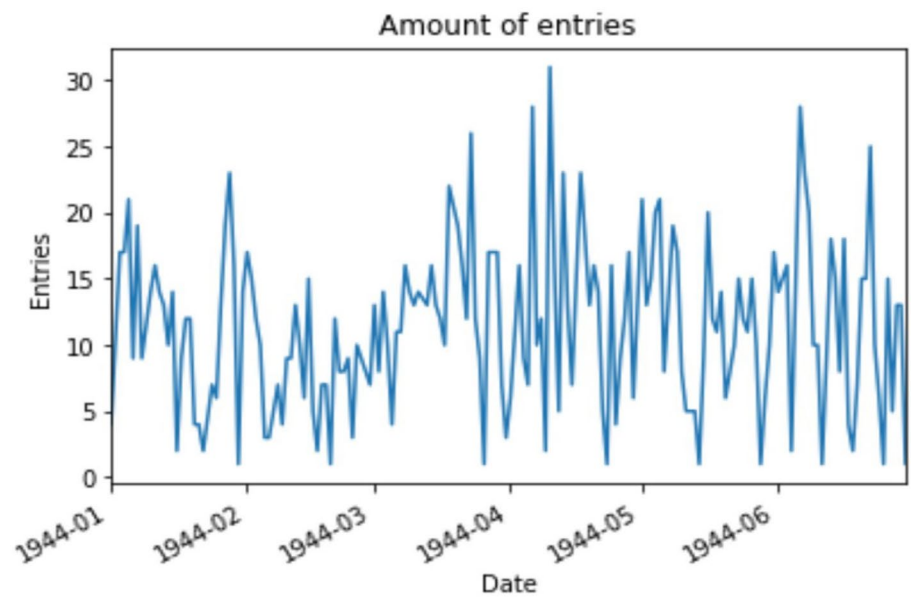


Fig. 12 Five most-associated words with “Oswego” over the first 60 volumes of 1944 (first 6 months)

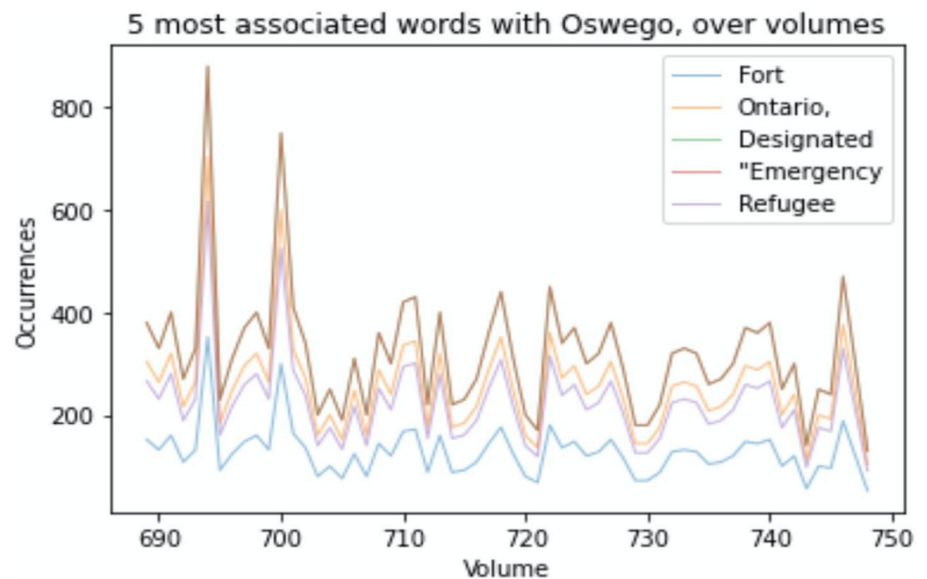
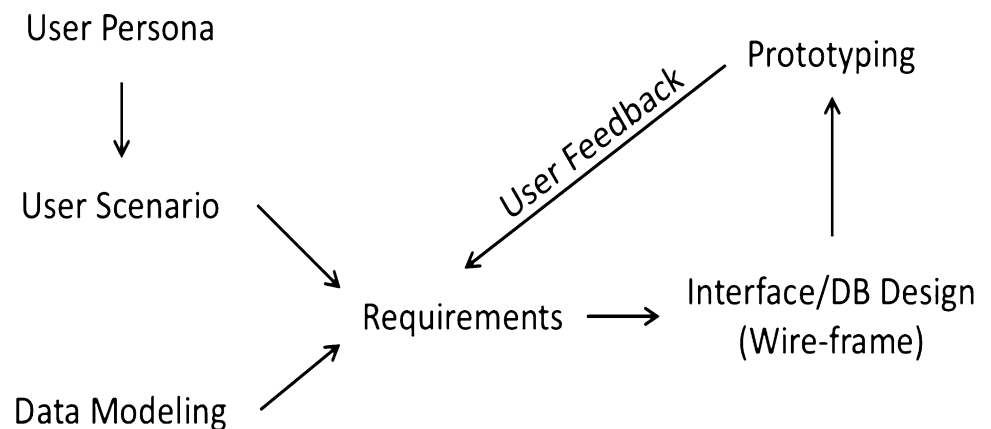


Fig. 13 Iterative design process used in the Human Face of Big Data (Lee et al. 2017)



```
# Query Volume 696 to find out the number of Subject Index entries.
entries = Query().withBook([696]).run()
numEntries = len(entries.content())
print("Number of entries: " + str(numEntries))
```

```
Number of entries: 31
```

```
# Query Volume 696 to find the list of entry headers.
entries = Query().withBook([696]).run()
entries.headers()
```

```
['Baerwald, Paul', 'Armstrong, Hamilton Fish', 'Bolivia',
'Argentina', 'Bruere, Henry', 'Bullitt, William C.',
'Financing, Government', 'Business Conditions',
'Connecticut, Bridgeport', 'Ernst, Morris', 'China', 'Lend-Lease',
'Murray, Philip', 'Mann, James', 'National Gallery of Art',
'Jeffers, William', 'Latin America', 'Dakar', 'France',
'Bridgeport, Connecticut', 'Pehle, John W.',
'Foreign Funds Control', 'Refugees', 'Revenue Revision',
'Taxation', 'Tollaksen, Leslie B. - Lieutenant Commander, USCGR',
'U.S.S.R.', 'War Refugee Board', 'Widener Art Collection',
'Willkie, Wendell L.', 'Wise, Rabbi Stephen S.']
```

```
# Show the 27th header entry
entries.headers()[27]
```

```
War Refugee Board
```

Similarly, we can query content, indexes, dates, and keywords for a particular entry. We can also retrieve ranges of entries across Volumes such as in:

considering that the Morgenthau Diaries document the daily activities of *Treasury* leadership, a government department most readers would not necessarily expect to be heavily

```
entries = Query().withBookBetween(689,748).run()
```

More sophisticated uses of the library (code not shown here, but to be published on the CASES portal) lead to the creation of graphs that summarize entries over the first 6 months of 1944 and the 60 Volumes in our test database. These graphs were created based on a series of queries developed by Gondek to simulate potential future queries by public users. Figure 10 is a word frequency query and answers the questions: *What are the most commonly occurring topics? Who and what are the most commonly mentioned people, organizations, and countries?* These three entity types are central to analysis of historical social networks and thematic frameworks (Gondek 2018). In qualitative data analysis, this is part of a “grounded theory methodology” in which theories are designed based on first analyzing data, rather than starting with theory (Strauss and Corbin 1998).

The results of the sample word frequency query reveal that the War Refugee Board was the second most commonly mentioned subject heading in the Morgenthau Diaries TOCs for the first 6 months of 1944. This is significant,

engaged in refugee matters. In fact, the War Refugee Board emerged through lobbying by Treasury staff including John Pehle, originally the head of Foreign Funds Control who became the Board’s first Executive Director. The majority of Board staff came from Treasury and the WRB offices were based in the Treasury building (USHMM). Figure 10 is also significant, because it shows which countries or world regions were the most frequently mentioned including Latin America, China, USSR, Argentina, and France. In April and May 1944, the U.S. government authorized American consular officers in Switzerland, Spain, and Portugal to issue up to 4000 (Switzerland) and 1000 (Spain and Portugal) visas to refugee children arriving from France during the first 6 months of 1944. The U.S. asked countries in Latin America and the Caribbean to accept these refugee children from France (Gondek 2020b, Hull 1944) (Fig. 11).

Figure 12 emerged as a response to Gondek’s research into the development of the Emergency Refugee Shelter at Fort Ontario, NY, established in June 1944 through months of concerted effort by staff of the War Refugee

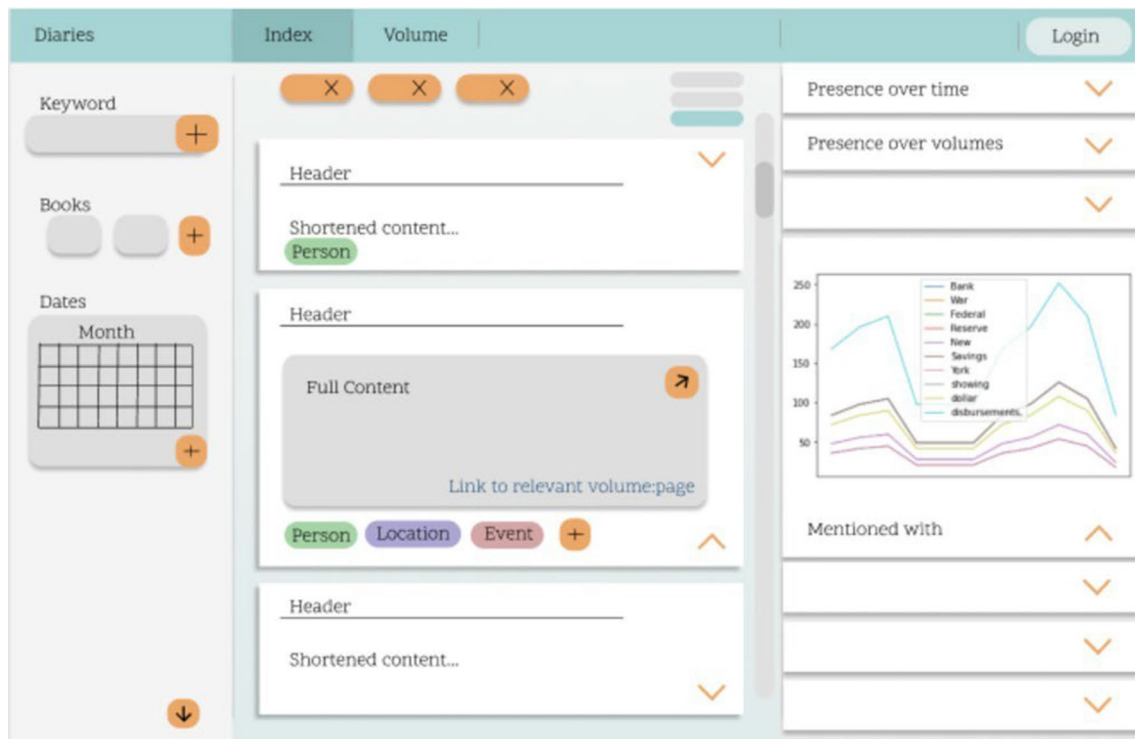


Fig. 14 Preliminary design of a new user interface that allows querying across many dimensions

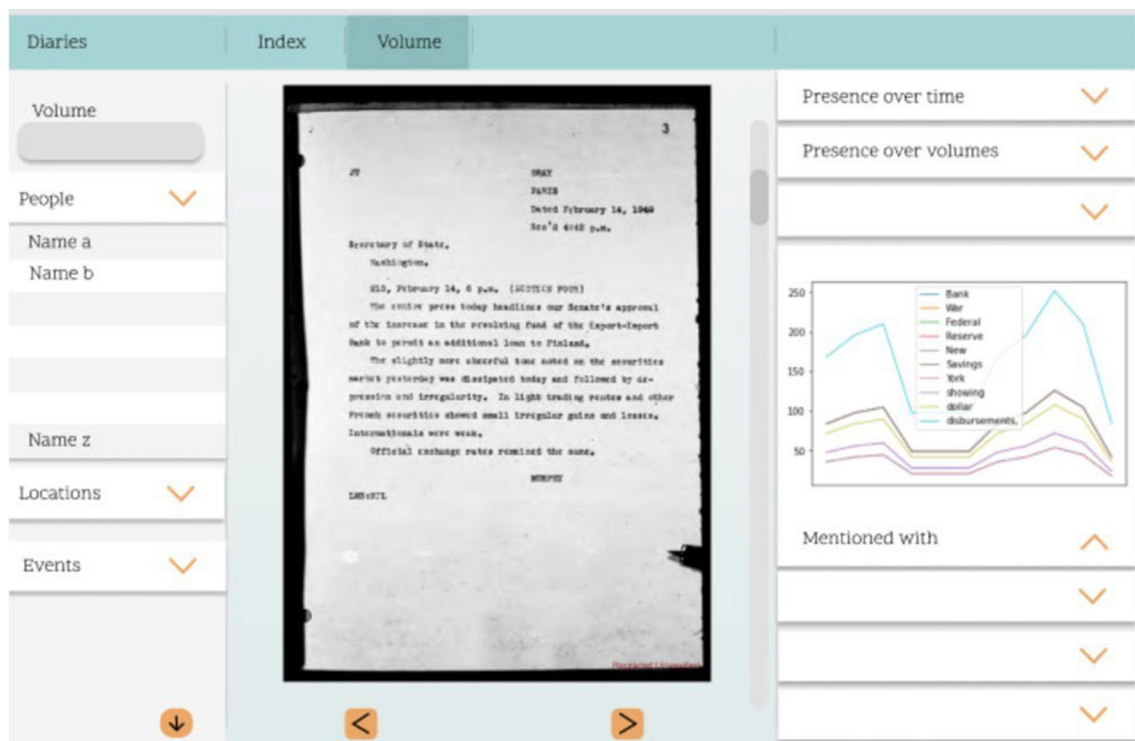


Fig. 15 The user interface will index and display Diary contents down to the document level

Board (Gondek 2021; Roosevelt 1944b). Refugees from Italy arrived at Fort Ontario in August 1944 (Myer 1944). The sample query used was: *which words were most commonly co-occurring with the term “Oswego?”* In May 1944, overpopulation of refugees in Southern Italy required that these refugees be moved elsewhere, and John Pehle and his staff at the WRB convinced the President to establish an Emergency Refugee Shelter at Fort Ontario, Oswego, New York, to house 1000 refugees (actual number rescued was 982) until the termination of the war. The WRB staff drafted earlier proposals for such a camp in March 1944 (Gondek 2021; Pehle 1944a, b, c). The purpose of creating this refugee camp within the U.S. was as a “token to the rest of the world that we, the United States Government, aren’t high and mighty in asking the rest of the world to do something which we aren’t willing to do ourselves” (Morgenthau Jr. 1944).

While this package shows the potential for telling computational stories with the Morgenthau Diaries and sharing these through interactive Jupyter Notebooks and as such represents a major breakthrough, this information display remains rather technical and specialized and will not necessarily be suited to all users in the general public. The next logical step is to build a more general-purpose user interface that could present alternative data formats and exploratory tools to non-expert users online.

5 Discussion

To encourage augmented exploration of primary sources by all kinds of users, this interdisciplinary collaboration identified the need for real-world outcomes in the form of improved online access to the Morgenthau Diaries. This goal, of course, is shared by the MHCP which strives to foster greater public understanding of American responses to the Holocaust. Since the MHCP has always planned to develop a new general-purpose user interface (an information portal to be named in honor of MHCP donor Peter S. Kalikow to debut in 2021), the Project embraced an iterative design strategy to assist in its development, primarily based on incorporating the computational concepts and treatments described in this paper. Ultimately, this interface must support the exchange of open data for digital humanists and provide improved content usability and curatorial tools for the general public.

5.1 Interface design to enable new modes of access

Our interdisciplinary collaborative team is emphasizing an iterative design process:

We leveraged the experimentations described so far in this paper by wireframing and prototyping an innovative

user interface for the MHCP to be published on the FDR Presidential Library’s website in October of 2021. The prototype illustrated in Fig. 13 displays sample query results based on those described in Sect. 4.3. The interface must assist users in identifying key themes and entities such as persons, organizations, locations, events, across volumes, and over time. A simple tool that queries the validated global index populated by ML object detection can aid in this goal. Users will be able to find when topics such as: “refugees,” “refugee camps,” “temporary haven,” “free port,” “Emergency Refugee Shelter,” “Oswego”, or “Fort Ontario” are mentioned, and further interrogate the records to find, for instance, *which people, organizations, countries/locations, and other topics are most frequently co-occurring? Who—including individuals and organizations—was most involved in the issue of refugee rescue and relief? Which volumes contain the highest number of mentions of these terms? Which time period?* Users also need the ability to navigate, directly from the descriptive references to the actual document in the Morgenthau Diaries (Figs. 14, 15).

5.2 Relevance of data and its structures to our users

If the Morgenthau Diaries are under-utilized by scholars and the general public, the problem is due to access and usability barriers and not to inherent research value. As described in Sect. 2, the Diaries represent an unusually valuable textual corpus for digital exploration for many reasons, but especially by virtue of its granular historical detail and the original finding tools baked into the source itself. While that complexity certainly adds value, it also presents serious challenges to functional access outside of the original, manual, in-person research environment for which its access schema were designed. New machine-aided finding tools can and must reduce those barriers.

This interdisciplinary team interrogated and analyzed important factors of the corpus, including its original and imposed descriptive data structures, relational context among record units, historical functional analysis, and digital versioning limitations. This was done all while centering community user needs as a lens for improved data access. Central questions driving both the data augmentation and conceptual interface design became *how can machine learning and artificial intelligence free data from restrictive or hidden settings while conveying and amplifying historical context for human access and understanding?* The team recognized the following as general user perspectives and user needs.

Users of augmented data produced by this collaboration may be:

- Traditional historical scholars with extensive subject knowledge from secondary and other primary sources,

seeking information ready for human reader qualitative analysis but appreciating computer-assisted finding.

- Digital humanities or sociological research scholars with extensive subject knowledge from secondary sources and other primary sources, seeking data already prepared for computational quantitative analysis.
- Students with varying levels of research experience, and varying levels of clarity in search or analysis goals.
- Reference archivists with broad knowledge of the collection (both context and content), and desire to improve and facilitate researcher use of a corpus. This role includes providing user instruction for interfacing with data, and performing preliminary content surveys in response to specific reference questions on demand.
- General public with a broad and perhaps passing interest.
- General public with a hyper-specific and persistent subject interest (family research especially).

Users may expect to:

- Search and find item-level source material related to key topics (subjects) using natural language terms.
- Immediately understand the relevance and relational context of specific search results.
- Immediately understand the formatting options for, and applications of all items returned by search and browse functions.
- Find information displayed at a level matching their individual interests and intentions for use, while (possibly) assuming their expectations are universal.

Conceptual suggestions for first iterations of MHCP public user interface:

- Present newly developed master faceted index for the general public to navigate and search across Volumes by subject.
- Present augmented datasets as open data available for new computational uses including Jupyter notebook, CSV, API, or other formats for digital humanists and data scientists who will perform their own digital analysis outside the MHCP interface.
- Describe and demonstrate possibilities for finding and display. Provide examples of digital scholarship output and/or provide embedded tools for visualizing information in a dataset.
- Either display results first in archival context within a human-readable hierarchy, or include archival context with all finding results.

It is important to note that the experimental way forward does not end or even culminate with the interface; rather, the public use of data should be concurrent with, and contribute

to, ongoing research and development in computational thinking. The limits of this case approach should be considered, as well, since its focus was to augment an original indexing framework for usability and interpretation by the public, and it does not use AI to generate new intellectual, interpretive topic models. Rather, the MHCP interface will become a living, growing, tangible product, one that invites public exploration and scholarly conversation as ML and AI experiments continue (Fig. 12). It is a new and real-world resource, one that the Library acknowledges may hold imperfect or evolving data, but one that provides and sustains an unprecedented level of access to key primary sources.

6 Conclusion

6.1 Benefits of interdisciplinary collaboration

This project relied heavily on combined regular input from highly skilled data and information scientists, an archivist deeply familiar with the corpus structure and its potential users, and a scholar with direct experience deriving and interpreting meaning from the corpus' content. Multiple perspectives and diverse professional expertise helped to balance the application of computational concepts throughout the experimentation period. Collaborative approaches to evaluating and responding to user needs allowed us to learn from one another and to incorporate those user needs into all aspects of data planning even prior to interface design.

Such early and thorough efforts to integrate perspectives and methods are uncommon in both archival practice and data science. The success we have enjoyed in this case shows proof-of-concept in key areas of augmented information strategy and overall applicability of collections as data. From an institutional perspective, neither the MHCP nor the AIC operating independently from one another could achieve the inclusive, data and user-driven outcomes demonstrated thus far. Future sustained collaborative efforts are imperative for moving forward with this progressive information strategy.

Part of the technological innovations that have emerged from our interdisciplinary collaboration are new culturally driven applications of ML and AI tools that have the potential to be expanded to broader classes of historical collections. Typical uses of Object Detection models are for drawing boxes around apples, flowers, or cars in images, and nanonets.com for capturing content from images of invoices, receipts, and driver's licenses. We believe our demonstration of new uses of these kinds of industry tools toward valuable historical and cultural analysis to be both innovative and original.

6.2 Future directions, scalability, and replicability

The MHCP and AIC will continue to collaborate on further experimental treatments and will contribute findings and information products to the evolving MHCP user interface. Additional innovative applications of ML, AI, and Digital Curation will be explored across these and related collections. There is a lack of experimentation in this field and a need to encourage sustained interdisciplinary collaborations.

Some specific, near-term goals are to extend data validation using the *Morgenthau Verifier* tool beyond the 6 month window of 1944 selected for the first computational proof of concept; incorporate crowdsourcing tools in the next iteration of the MHCP public interface; and to investigate additional knowledge organization systems and linked data methods for connecting digitized source images to one another through emerging facets of structured data.

Reproducibility to broad classes of collections was identified as important from the outset. We believe that this case study contributes substantially to revealing the strengths and weaknesses of using AI/ML systems in cultural organizations, particularly with regard to adapting original indexing schema. These treatments show computational methods applied to a large textual corpus, show object detection can accomplish goals usually attempted with named entity recognition, and show that it is possible to center user needs throughout the information strategy and experimentation process. These processes can be applied and customized to fit any type of textual collection, not only those with complex legacy classification included. Though most historical archival collections do not include item-level indices contemporary to the records they describe, there are actually several examples of such sources throughout the FDR Library's archives. This is partly due to the age of our repository (collections were first processed and opened during and immediately following the Roosevelt Administration, when card indices were used often as manual access tools). FDR's Papers as President and Eleanor Roosevelt's Papers as First Lady came to the archives directly from the White House, and like the Morgenthau Diaries, the original order and composite record keeping systems were kept intact. Those collections include complex numerical filing schema, cross-references embedded through file notes and/or associated correspondence indices.

The FDR Library looks forward to testing these computational treatments at a larger scale, replicating these and additional ML and AI methods on much larger collections once those sources are digitized or datafied, though no funding is currently in place to support specific projects at this time. First in line will be the Eleanor Roosevelt Papers, a 3 million-page textual collection documenting her incredible life and work, a corpus that comes complete with an

historical, contemporary item-level index to her White House correspondence.

6.3 Lessons learned

This team recognized from the outset that to improve access to the Morgenthau Diaries successfully, we had to be flexible and accept the nature of this collaboration as ongoing, iterative, and likely to include setbacks. One example of trial and error involved attempting treatments on the index cards first, then pivoting to the TOCs as their full-page information structure proved more consistent with fewer handwriting barriers. This led to more in-depth understanding on our part, both computationally and comprehensively, about the nature and function of the source material itself. Another example occurred during the query sampling process. After several group attempts to our research scholar determined that without additional support and mediation through a user interface, she most likely would not be able to use or interpret results returned. Rather than set us back, we considered the finding as evidence to support and demonstrate the need for user-centered interface design.

We adopted a similar attitude toward conceptual interface development, confident in our belief that the public will appreciate information shared *throughout* the process of data augmentation rather than waiting for a more perfect or finished system. This represents a pragmatic approach that may seem risky to some institutions, but is entirely consistent with the path-finding goals set forth by the MHCP, AIC, and by the FDR Library in all of its digital access initiatives inspired by Franklin D. Roosevelt's characteristic call for "bold, persistent experimentation" (Roosevelt 1932). The collaborative nature of our work led to a deeper understanding of the value of inclusive community involvement and drives future directions for the interface, for experimentation to further reveal and disseminate data, and to increase public engagement with primary sources, especially on complex and controversial areas such as the Holocaust.

Acknowledgements We wish to acknowledge funding from the 2020–2022 IMLS Laura Bush 21st Century Librarian Program: "Piloting an Online Collaborative Network for Integrating Computational Thinking into Library and Archival Education and Practice." This grant supports the piloting of an online collaborative network of educators and practitioners to enable the sharing and dissemination of computational case studies and lesson plans through a Jupyter Notebook interactive computational learning platform, called CASES [Computational Archival Science Educational System], see: <https://cases.umd.edu>. We also hope to share the lessons learned from exploring the Morgenthau collections. We also wish to acknowledge funding from the Morgenthau family, generous friends of the family, and grants from the Righteous Persons Foundation and the Krupp family that have supported the Morgenthau Holocaust Collections Project since 2018. These contributions fund the vital research of the Morgenthau Scholar, the upcoming international conference and exhibit, as well as the creation of new, essential digital infrastructure for the study of the Holocaust. Supplementary funding

came from the New York Community Trust which made the *Investigating the Holocaust* video series and curriculum resources possible. <https://support.rooseveltinstitute.org/campaign/library-mhpc/c243440>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adalian RP (2019) Morgenthau, Ambassador Henry, Sr. Encyclopedia entries on the Armenian Genocide. Armenian National Institute. <https://www.armenian-genocide.org/morgenthau.html>. Accessed 27 Apr 2021
- Beschloss M (2002) *The conquerors: Roosevelt, Truman, and the destruction of Hitler's Germany, 1941–1945*. Simon & Schuster, New York
- Breitman R, Lichtman A (2013) *FDR and the Jews*. Belknap Press, Cambridge
- Chodorow K, Dirolf M (2010) *MongoDB: the definitive guide*, 3rd edn. O'Reilly Media, Newton
- Colavizza G, Blanke T, Jeurgens C, Noordegraaf J (2021) Archives and AI: an overview of current debates and future perspectives. *J Comput Cult Herit (JOCCH 20-0191)* Spec Issue Comput Arch Sci
- Cordell R (2020) Machine learning + libraries, a report on the state of the field, commissioned by LC Labs, Library of Congress. <https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf>
- Diamond I (1941) Treasury Department Inter-Office Communication from Miss Isabella Diamond to Mrs. Henrietta Klotz, 2 August 1941. Di thru Dn, 1940–1942 Folder, Morgenthau Correspondence, Di-Dn [1933]—Diary of HM Jr. [1933–1944] (Corres. Re) Box 77, Morgenthau Papers, FDR Presidential Library and Museum, Hyde Park, NY
- Diamond I (1943) Memorandum from Miss Isabella Diamond to Mrs. Henrietta Klotz, 17 December 1943. Di thru Dn, 1943–1945 Folder, Morgenthau Correspondence, Di-Dn [1933]—Diary of HM Jr. [1933–1944] (Corres. Re) Box 77, Morgenthau Papers, FDR Presidential Library and Museum, Hyde Park, NY
- Erbelding R (2015) Morgenthau family papers, 1860–2015, 2015.255.1 Finding Aid. United States Holocaust Memorial Museum, Washington D.C. <https://collections.ushmm.org/search/catalog/irn96059>
- Erbelding R (2018) *Rescue board: the untold story of America's efforts to save the Jews of Europe*. Doubleday, New York
- FDR Presidential Library and Museum (2018a) Finding aid diaries of Henry Morgenthau Jr., April 27, 1933–July 27, 1945. FDR Presidential Library and Museum. <http://www.fdrlibrary.marist.edu/archives/collections/franklin/index.php?p=collections/findingaid&id=535>. Accessed 27 Apr 2021
- FDR Presidential Library and Museum (2018b) Finding aid Henry Morgenthau Jr. papers, 1866–1953. FDR Presidential Library and Museum. <http://www.fdrlibrary.marist.edu/archives/collections/franklin/index.php?p=collections/findingaid&id=159&q=&rootentid=72431>. Accessed 27 Apr 2021
- FDR Presidential Library and Museum (2021) Morgenthau Holocaust collections project. <https://www.fdrlibrary.org/morgenthauproject>
- Gondek A (2018) Jewish women's transracial epistemological networks: representations of black women in the African diaspora, 1930–1980. Dissertation, Florida International University
- Gondek A (2020a) Hidden figures—Henrietta Stein Klotz: “The watchdog of the Secretary of the Treasury.” Forward with Roosevelt. FDR Presidential Library and Museum. <https://fdr.blogs.archives.gov/2020/03/24/updates-morgenthau-holocaust-collections-project-2/>. Published 24 March 2020. Accessed 27 Apr 2021
- Gondek A (2020b) Jewish refugee children and the establishment of the War Refugee Board, 1943–1944, a path through the Morgenthau Diaries and the War Refugee Board Papers. <https://abbysgondek.com/portfolio/jewish-refugee-children-and-the-establishment-of-the-war-refugee-board-1943-1944-a-path-through-the-morgenthau-diaries-and-war-refugee-board-papers/>. Accessed 27 Apr 2021
- Gondek A (2021) Letters from the public in support of a refugee camp at Oswego, NY (1944). Adobe Spark. <https://spark.adobe.com/page/xsX0z3D9IMyq1/>. Accessed 27 Apr 2021
- Hull C (1944) Cordell Hull to ambassadors at Panama, Habana, Ciudad Trujillo, Bogota, Lima, Santiago, Montevideo and Mexico, D.F., Circular Airgram. WRB Series 10, Box 117, History of War Refugee Board with Selected Documents, vol II, Folder 2. FDR Presidential Library and Museum, Hyde Park, p 570
- Klotz H (1986) Letter to Henry Morgenthau III from Herman Klotz, 16 October 1986. Morgenthau Family Papers 2015.255.1, Box 32, File 11, Henrietta Klotz File, USHMM, Washington, DC
- Lee M, Zhang Y, Chen S, Spencer E, Dela Cruz J, Hong H, Marciano R (2017) Heuristics for assessing computational archival science (CAS) research: the case of the human face of big data project. The workshop on computational archival science in IEEE Big Data. Boston, US, pp 2262–2270. https://ai-collaboratory.net/wp-content/uploads/2020/04/Myeong_Lee.pdf
- Marciano R, Jansen G, Underwood W (2019) Developing a framework to enable collaboration in computational archival science education. SAA 2019 Research Forum
- McReynolds (1939) Memorandum from Mr. McReynolds to Miss Isabella Diamond, 3 February 1939. Isabella Diamond Folder, Morgenthau Correspondence, Di-Dn [1933]—Diary of HM Jr. [1933–1944] (Corres. Re) Box 77, Morgenthau Papers, FDR Presidential Library and Museum, Hyde Park, NY
- Morgenthau Jr H (1936) Letter from Henry Morgenthau Jr. to Isabella Diamond, 17 August 1936. Di thru Dn 1933–1936 Folder, Morgenthau Correspondence, Di-Dn [1933]—Diary of HM Jr. [1933–1944] (Corres. Re) Box 77, Morgenthau Papers, FDR Presidential Library and Museum, Hyde Park, NY
- Morgenthau Jr H (1944) Meeting transcript between HMJ, John Pehle, Josiah Du Bois and Henrietta Klotz to discuss the conversation with Mr. McCloy. Morgenthau Diaries 738:179–180. FDR Presidential Library and Museum, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/morg/md1025.pdf
- Morgenthau Jr H (1945) Handwritten letter from Henry Morgenthau Jr. to Mrs. Henrietta Klotz, 5 August 1945. Morgenthau Family Papers 2015.255.1, Box 32, File 11, Henrietta Klotz File, USHMM, Washington, DC
- Morgenthau H III (1991) *Mostly Morgenthau: a family history*. Ticknor & Fields, New York
- Morgenthau H Jr, Pehle J, DuBois J (1944) Jewish evacuation meeting transcript. Morgenthau Diaries 707:219–233
- Morgenthau Family Papers, specific author unknown (n.d.) Undated Memo to Explain Henry Morgenthau Jr.'s August 5, 1945 Letter to Henrietta Klotz. Morgenthau Family Papers Accession

- 2015.255.1, Box 32, Folder 11 Henrietta Klotz: pages 184–185 in file. United States Holocaust Memorial Museum, Washington, DC
- Myer D (1944) Report on emergency refugee shelter. Morgenthau Diaries 779:162–170. FDR Presidential Library, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/morg/md1073.pdf
- New York Times (NYT) Archive (1988) H.S. Klotz, 87, Aide to Treasury Secretary. New York Times Online. 21 December. <https://www.nytimes.com/1988/12/21/obituaries/h-s-klotz-87-aide-to-treasury-secretary.html>
- Padilla T, Allen L, Frost H, Potvin S, Russey Roke E, Varner S (2019) Final report—always already computational: collections as data. Zenodo. <https://doi.org/10.5281/zenodo.3152935>
- Paul R (1944) Report to the secretary of the acquiescence of this Government in the murder of the Jews. Morgenthau Diaries 693:212–229. FDR Presidential Library, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/morg/md0978.pdf
- Pehle J (1944a) Report to the War Refugee Board. Morgenthau Diaries 707:235–240. FDR Presidential Library, Hyde Park, NY
- Pehle J (1944b) Memorandum for the President. Morgenthau Diaries 716:171–174. FDR Presidential Library, Hyde Park, NY
- Pehle J (1944c) Memorandum for the President. Morgenthau Diaries 726:40–47. FDR Presidential Library, Hyde Park, NY
- Penkower MN (2016) The Earl Harrison report: its genesis and its significance. *Am Jew Arch J* LXVIII(1):1–75
- Perkel J (2018) Why Jupyter is data scientists' computational notebook of choice. *Nature* 563:145–146 (2018). See: <https://www.nature.com/articles/d41586-018-07196-1>
- Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: a python natural language processing toolkit for many human languages. In: Association for computational linguistics (ACL) system demonstrations [pdf]
- Randby T, Marciano R (2020) Digital curation and machine learning experimentation in archives. In: Computational archival science workshop #5, proceedings 2020 IEEE international conference on big data, 11 December
- Roosevelt FD (1932) Atlanta, Georgia: Oglethorpe University Commencement Address (speech file 476). FDR's Papers as President, Master Speech File. FDR Presidential Library, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/msf/msf00486
- Roosevelt FD (1944a) Executive order 9417: establishing a War Refugee Board. Henry Morgenthau Jr. Diaries 696:1–3. FDR Presidential Library, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/morg/md0981.pdf
- Roosevelt FD (1944b) To the Congress of the United States. Morgenthau Diaries 742:296–298. FDR Presidential Library, Hyde Park, NY. http://www.fdrlibrary.marist.edu/_resources/images/morg/md1030.pdf
- Strauss A, Corbin J (1998) Basics of qualitative research: techniques and procedures for developing grounded theory, 2nd edn. Sage Publications, Thousand Oaks
- Tai J, Zavala J, Gabiola J, Brilmyer G, Caswell M (2019) Summoning the ghosts: records as agents in community archives. *J Contemp Arch Stud* 6, Art. 18:1–20. <https://elischolar.library.yale.edu/jcas/vol6/iss1/18>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.