DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 4

Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset

Benjamin Lee <bcgl_at_cs_dot_washington_dot_edu>, The Library of Congress & The University of Washington

Abstract

The increasing roles of machine learning and artificial intelligence in the construction of cultural heritage and humanities datasets necessitate critical examination of the myriad biases introduced by machines, algorithms, and the humans who build and deploy them. From image classification to optical character recognition, the effects of decisions ostensibly made by machines compound through the digitization pipeline and redouble in each step, mediating our interactions with digitally-rendered artifacts through the search and discovery process. As a result, scholars within the digital humanities community have begun advocating for the proper contextualization of cultural heritage datasets within the socio-technical systems in which they are created and utilized. One such approach to this contextualization is the *data archaeology*, a form of humanistic excavation of a dataset that Paul Fyfe defines as "recover[ing] and reconstitut[ing] media objects within their changing ecologies" [Fyfe 2016]. Within critical data studies, this excavation of a dataset - including its construction and mediation via machine learning - has proven to be a capacious approach. However, the data archaeology has yet to be adopted as standard practice among cultural heritage practitioners who produce such datasets with machine learning.

In this article, I present a data archaeology of the Library of Congress's *Newspaper Navigator* dataset, which I created as part of the Library of Congress's Innovator in Residence program [Lee et al. 2020]. The dataset consists of visual content extracted from 16 million historic newspaper pages in the *Chronicling America* database using machine learning techniques. In this case study, I examine the manifold ways in which a *Chronicling America* newspaper page is transmuted and decontextualized during its journey from a physical artifact to a series of probabilistic photographs, illustrations, maps, comics, cartoons, headlines, and advertisements in the *Newspaper Navigator* dataset [Fyfe 2016]. Accordingly, I draw from fields of scholarship including media archaeology, critical data studies, science and technology studies, and the autoethnography throughout.

To excavate the *Newspaper Navigator* dataset, I consider the digitization journeys of four different pages in Black newspapers included in *Chronicling America*, all of which reproduce the same photograph of W.E.B. Du Bois in an article announcing the launch of *The Crisis*, the official magazine of the NAACP. In tracing the newspaper pages' journeys, I unpack how each step in the *Chronicling America* and *Newspaper Navigator* pipelines, such as the imaging process and the construction of training data, not only imprints bias on the resulting *Newspaper Navigator* dataset but also propagates the bias through the pipeline via the machine learning algorithms employed. Along the way, I investigate the limitations of the *Newspaper Navigator* dataset and machine learning techniques more generally as they relate to cultural heritage, with a particular focus on marginalization and erasure via algorithmic bias, which implicitly rewrites the archive itself.

In presenting this case study, I argue for the value of the data archaeology as a mechanism for contextualizing and critically examining cultural heritage datasets within the communities that create, release, and utilize them. I offer this autoethnographic investigation of the *Newspaper Navigator* dataset in the hope that it will be considered not only by users of this dataset in particular but also by digital humanities practitioners and end users of cultural heritage datasets writ large.

I. An Introduction to the Newspaper Navigator Dataset

1

In partnership with LC Labs, the National Digital Newspaper Program, and IT Design & Development at the Library of Congress, as well as Professor Daniel Weld at the University of Washington, I constructed the *Newspaper Navigator* dataset as the first phase of my Library of Congress Innovator in Residence project, *Newspaper Navigator*.[1] The project has its origins in *Chronicling America*, a database of digitized historic American newspapers created and maintained by the National Digital Newspaper Program, itself a partnership between the Library of Congress and the National Endowment for the Humanities. Content in *Chronicling America*is contributed by state partners of the National Digital Newspaper Program who have applied for and received awards from the Division of Preservation and Access at the National Endowment for the Humanities [Mears 2014]. At the time of the construction of the *Newspaper Navigator* dataset in March, 2020, *Chronicling America* contained approximately 16.3 million digitized historic newspaper pages published between 1789 and 1963, covering 47 states as well as Washington, D.C. and Puerto Rico. The technical specifications of the National Digital Newspaper Program require that each digitized page in *Chronicling America* comprises the following digital artifacts [National Digital Newspaper Program 2020]:

1. A page image in two raster formats:
    1. Grayscale, scanned for maximum resolution possible between 300-400 DPI, relative to the original material, uncompressed TIFF 6.0
    2. Same image, compressed as JPEG2000
2. Optical character recognition (OCR) text and associated bounding boxes for words (one file per page image)
3. PDF Image with Hidden Text, i.e., with text and image correlated
4. Structural metadata (a) to relate pages to title, date, and edition; (b) to sequence pages within issue or section; and (c) to identify associated image and OCR files
5. Technical metadata to support the functions of a trusted repository

2

Additional artifacts and metadata are contributed for each digitized newspaper issue and microfilm reel. All digitized pages are in the public domain and are available online via a public search user interface,[2] making *Chronicling America* an immensely rich resource for the American public.

3

The central goal of *Newspaper Navigator* is to re-imagine how the American public explores *Chronicling America* by utilizing emerging machine learning techniques to extract, categorize, and search over the visual content and headlines in *Chronicling America*'s 16.3 million pages of digitized historic newspapers. *Newspaper Navigator* was both inspired and directly enabled by the Library of Congress's *Beyond Words* crowdsourcing initiative [Ferriter 2017]. Launched by LC Labs in 2017, *Beyond Words* engages the American public by asking volunteers to identify and draw boxes around photographs, illustrations, maps, comics, and editorial cartoons on World War I-era pages in *Chronicling America*, note the visual content categories, and transcribe the relevant textual information such as titles and captions.[3] The thousands of annotations created by *Beyond Words* volunteers are in the public domain and available for download online. *Newspaper Navigator* directly builds on *Beyond Words* by utilizing these annotations, as well as additional annotations of headlines and advertisements, to train a machine learning model to detect visual content in historic newspapers.[4] Because *Beyond Words* volunteers were asked to draw bounding boxes to include any relevant textual content, such as a photograph's title, this machine learning model learns during training to include relevant textual content when predicting bounding boxes.[5] Furthermore, in the *Transcribe* step of *Beyond Words*, the system provided the OCR with each bounding box as an initial transcription for the volunteer to correct; inspired by this, the *Newspaper Navigator* pipeline automatedly extracts the OCR falling within each predicted bounding box in order to provide noisy textual metadata for each image. In the case of headlines, this method enables the headline text to be directly extracted from the bounding box predictions. Lastly, the pipeline generates image embeddings for the extracted visual content using an image classification model trained on ImageNet.[6] A diagram of the full *Newspaper Navigator* pipeline can be found in Figure 1.

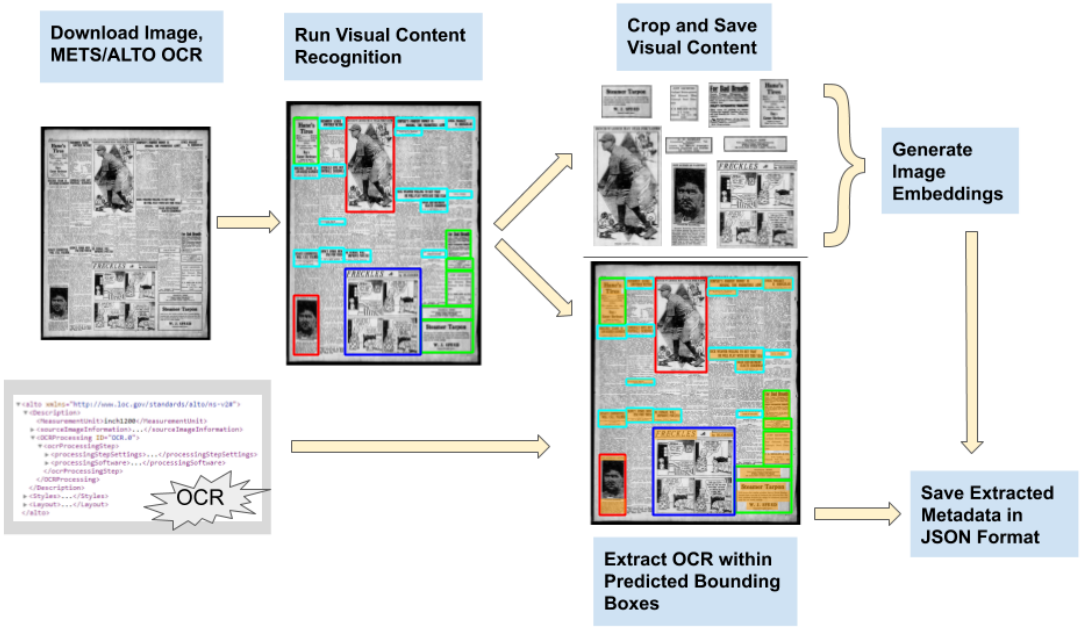

Figure 1. A diagram showing the *Newspaper Navigator* pipeline, which processed over 16.3 million historic newspaper pages in *Chronicling America*, resulting in the *Newspaper Navigator* dataset.

Over the course of 19 days from late March to early April of 2020, the *Newspaper Navigator* pipeline processed 16.3 million pages in *Chronicling America*; the resulting *Newspaper Navigator* dataset was publicly released in May, 2020. The full dataset, as well as all code written for this project, are available online and have been placed in the public domain for unrestricted re-use.[7] Currently, the *Newspaper Navigator* dataset can be queried using HTTPS and Amazon S3 requests. Furthermore, hundreds of pre-packaged datasets have been made available for download, along with associated metadata. These pre-packaged datasets consist of different types of visual content for each year, from 1850 to 1963, allowing users to download, for example, all of the maps from 1863 or all of the photographs from 1910. For more information on the technical aspects of the pipeline and the construction of the *Newspaper Navigator* dataset, I refer the reader to [Lee et al. 2020] 4

II. Why a Data Archaeology?

As machine learning and artificial intelligence play increasing roles in digitization and digital content stewardship, the Libraries, Archives, and Museums ("LAM") community has repeatedly emphasized the importance of ensuring that these emerging methodologies are incorporated ethically and responsibly. Indeed, a major theme that emerged from the "Machine Learning + Libraries Summit" hosted by LC Labs in September, 2019, was that "there is much more 'human' in machine learning than the name conveys" and that transparency and communication are first steps toward addressing the "human subjectivities, biases, and distortions" embedded within machine learning systems [LC Labs and Digital Strategy Directorate 2020]. This data archaeology has been written in support of this call for transparency and responsible stewardship, which is echoed in the Library of Congress's Digital Strategy, as well as the recommendations in Ryan Cordell's report to the Library of Congress *ML + Libraries: A Report on the State of the Field,* Thomas Padilla's OCLC position paper *Responsible Operations: Science, Machine Learning, and AI in Libraries*, and the University of Nebraska-Lincoln's report on machine learning to the Library of Congress [Library of Congress 2019]; [Cordell 2020]; [Padilla 2019]; [Lorang et al 2020]. I write this data archaeology from my perspective of having created the dataset, and although I am not without my own biases, I have attempted to represent my work as honestly as possible. Accordingly, I seek not only to document the construction of the *Newspaper Navigator* dataset through the lens of data stewardship but also to critically examine the dataset's limitations. In doing so, I advocate for the importance of autoethnographic approaches to documenting a cultural heritage dataset's construction from a humanistic perspective. 5

This article draws inspiration from recent works in media and data archaeology, including Paul Fyfe's "An Archaeology of Victorian Newspapers"; Bonnie Mak's "Archaeology of a Digitization"; Kate Crawford and Trevor Paglen's "Excavating 6

AI: The Politics of Images in Machine Learning Training Sets"; and, most directly, Ryan Cordell's "Qi-jtb the Raven: Taking Dirty OCR Seriously," in which Cordell traces the digitization of a single issue of the *Lewisburg Chronicle* from its selection by the Pennsylvania Digital Newspaper Project to its ingestion into the *Chronicling America* online database, with a focus on the distortive effects of OCR [Fyfe 2016]; [Mak 2017]; [Crawford and Paglen 2019]; [Cordell 2017]. As argued by Trevor Owens and Thomas Padilla, it is essential to "document how digitization practices and how the affordances of particular sources … produce unevenness in the discoverability and usability of collections" [Owens and Padilla 2020]. Recent works within the machine learning literature have analogously emphasized the importance of documenting the collection and curation efforts underpinning community datasets and machine learning models. Reporting mechanisms include "Datasheets for Datasets," "Dataset Nutrition Labels," "Data Statements for NLP," "Model Cards for Model Reporting," and "Algorithmic Impact Assessments" [Gebru et al. 2020]; [Holland et al. 2018]; [Bender and Friedman 2018]; [Mitchell et al. 2019]; [Reisman et al. 2018]. This case study adopts a similar framing in stressing the importance of reporting mechanisms, with a particular focus on the data archaeology in the context of cultural heritage datasets.

In the following sections, I trace the digitization process and data flow for *Newspaper Navigator*, beginning with the physical artifact of the newspaper itself and ending with the machine learning predictions that constitute the *Newspaper Navigator* dataset, reflecting on each step through the lens of discoverability and erasure. In particular, I study four different *Chronicling America* Black newspaper pages published in 1910, each depicting the same photograph of W.E.B. Du Bois, as the pages move through the *Chronicling America* and *Newspaper Navigator* pipelines. All four pages reproduce the same article by Franklin F. Johnson, a reporter from *The Baltimore Afro-American* [Farrar 1998]; the headline is as follows: 7

NEW MOVEMENT 8

BEGINS WORK 9

Plan and Scope of the Asso- 10

ciation Briefly Told. 11

Will Publish the Crisis. 12

Review of Causes Which Led to the 13

Organization of the Association in 14

New York and What Its Policy Will 15

Be-Career and Work of Professor 16

W.E.B. Du Bois 17

The article describes the creation of the National Association for the Advancement of Colored People (NAACP), details W.E.B. Du Bois's background, and announces the launch of *The Crisis*, the official magazine of the NAACP, with Du Bois as Editor-in-Chief. The four pages comprise the front page of the October 14th, 1910, issue of the *Iowa State Bystander* [Iowa State Bystander 1910]; the 16th page of the October 15th, 1910, issue of *Franklin's Paper the Statesman* [Franklin's Paper the Statesman 1910]; and the 2nd and 3rd pages of the October 15th, 1910, and November 26th, 1910, issues of *The Broad Ax*, respectively [The Broad Ax 1910a]; [The Broad Ax 1910b]. All four digitized pages are reproduced in the Appendix. 18

III. *Chronicling America*: A Genealogy of Collecting, Microfilming, and Digitizing

Any examination of *Newspaper Navigator* must begin with the genealogy of collecting, microfilming, and digitizing that dictates which newspapers have been ingested into the *Chronicling America* database. The question of what to digitize 19

is, in practice, answered and realized incrementally over decades, beginning at its most fundamental level with the question of which newspapers have survived and which have been reduced to lacunae in the historical record [Hardy and DiCuirci 2019]. [8] Historic newspapers present challenges for digitization in part due to the ephemerality of the physical printed newspaper itself: many newspapers were microfilmed and immediately discarded due to a fear that the physical pages would deteriorate.[9] Indeed, almost all of the pages included in *Chronicling America* have been digitized directly from microfilm. In the next section, I will examine the microfilm imaging process in more detail; however, in most cases, librarians selected newspapers for collecting and microfilming decades before the National Digital Newspaper Program was launched in 2004. These selections were informed by a range of factors including historical significance - itself a subjective, nebulous, and ever-evolving notion that has historically served as the basis for perpetuating oppression within the historical record. In "Chronicling White America," Benjamin Fagan highlights the paucity of Black newspapers in *Chronicling America*, in particular in relation to pre-Civil War era newspapers [Fagan 2016]. It is imperative to remember that this paucity can directly be traced back decades to the collecting and preserving stages.[10]

20 In regard to collecting, the newspaper page is both an informational object (i.e., the newspaper page as defined by its content) and a material object (i.e., the specific printed copy of the newspaper page) [Owens 2018]. At some point in time, librarians accessioned a specific copy of each printed page and microfilmed it or contracted out the microfilming. The materiality of that specific printed page is a confluence of unique ink smudges, rips, creases, and page alignment, much of which is captured in the microfilm imaging process. Though we may not make much of a crease or a smudge on a digitized page when we find it in the *Chronicling America* database, it can very well take on a life of its own with a machine learning algorithm in *Newspaper Navigator*. The machine learning algorithm might deem two newspaper photographs as similar simply due to the presence of creases or smudges, even if the photographs are easily discernible to the naked eye, or the smudges are of entirely different origin (i.e., a printing imperfection versus a smudge from a dirty hand).

21 It is only by foregrounding these subtleties of the collection, preservation, and microfilming processes that we can understand the selection process for *Chronicling America* in its proper context. The grant-seeking process dictates selection criteria for *Chronicling America* by which state-level institutions including state libraries, historical societies, and universities apply for two years of grant funding from the National Digital Newspaper Program via the Division of Preservation and Access at the National Endowment for the Humanities. With the awarding of a grant, a state-level awardee then digitizes approximately 100,000 newspaper pages published in their state for inclusion in *Chronicling America* [National Digital Newspaper Program 2020]; [NEH Division of Preservation and Access 2020]. The grant-seeking and awarding process is nuanced, but salient points include that state-level applicants must assemble an advisory board including scholars, teachers, librarians, and archivists to aid in the selection of newspapers, and grants are reviewed by National Endowment for the Humanities staff, as well as peer reviewers.[11]

22 Regarding selection criteria for newspaper titles, the National Digital Newspaper Program defines the following factors for state-level awardees to consider for content selection after a newspaper is determined to be in the public domain [National Digital Newspaper Program no date]:

- image quality in the selection of microfilm
- research value
- geographic representation
- temporal coverage
- bibliographic completeness of microfilm copy
- diversity (i.e., "newspaper titles that document a significant minority community at the state or regional level")
- whether the title is orphaned (i.e., whether the newspaper has "ceased publication and lack[s] active ownership" [Chronicling America no date])
- whether the title has already been digitized.

23 Though factors such as research value are considered by each state awardee's advisory board, as well as by the

National Endowment for the Humanities and peer review experts, the titles included in *Chronicling America* are largely dictated by which exist on microfilm and are of sufficient image quality within a state-level grantee's collection. Thus, the significance of the collection and microfilming practices of decades prior cannot be understated.

I also highlight that assessing microfilmed titles based on image quality is a complex procedure in its own right. The National Digital Newspaper Program has made publicly available a number of resources devoted specifically to this task, including documents and video tutorials [Barrall and Guenther 2005]; [Meta | Morphosis no date]. They articulate factors such as the microfilm generation (archive master, print master, or review copy), the material (polyester or acetate), the reduction ratio, and the physical condition. The detailed resources made available by the National Digital Newspaper Program, the Library of Congress, and the National Endowment for the Humanities for navigating this process are testaments to the multidimensional complexity of the selection process for *Chronicling America* [National Digital Newspaper Program 2019]; [National Digital Newspaper Program no date]; [NEH Division of Preservation and Access 2020]. 24

We have not yet investigated the topic of digitization, and we have already encountered a profusion of factors from collection to digitization that mediate which artifacts appear in *Chronicling America* and thus *Newspaper Navigator*. Let us now examine the microfilm itself. 25

IV. The Microfilm

In "What Computational Archival Science Can Learn from Art History and Material Culture Studies," Lyneise Williams shares a powerful anecdote of coming across a physical copy of a 1927 issue of the French sports newspaper *Match L'Intran* that featured accomplished Black Panamanian boxer, Alfonso Teofilo Brown, on the front cover [Williams 2019]. Williams describes Brown as "glowing. He looked like a 1920s film star rather than a boxer" [Williams 2019]. Curious to learn more about the printing process, Williams discovered that the issue of *Match L'Intran* was produced using rotogravure, a specific printing process that could "capture details in dark tones" [Williams 2019]. However, when Williams found a version of the same newspaper cover that had been digitized from microfilm, it was apparent that the microfilming process had washed out the detail of the rotogravure, reducing Brown to a "flat black, cartoonish form" [Williams 2019]. Williams relays the anecdote to articulate that the microfilming process itself is thus a form of erasure for communities of color [Williams 2019]. 26

The grayscale saturation of photographs induced by microfilming is widely documented and recognizable to most researchers who have ever worked with the medium [Baker 2001]; however, Lyneise Williams's article affords us a lens into what precisely is lost amongst the distortive effects of the microfilming process. This erasure via microfilming can be seen in *Chronicling America* directly. In Figure 2, I show the same photograph of W.E.B. Du Bois as it appears in 4 different *Chronicling America* newspaper pages published during October and November of 1910 and digitized from microfilm [Iowa State Bystander 1910]; [Franklin's Paper the Statesman 1910]; [The Broad Ax 1910a]; [The Broad Ax 1910b]. The phenomenon described by Williams is immediately recognizable in these four images: Du Bois's facial features are distorted by the grayscale saturation. In the case of the *Iowa State Bystander*, Du Bois has been rendered into a silhouette. 27

Moreover, each digitized reproduction reveals unique visual qualities, varying in contrast, sharpness, and noise - a testament to the confluence of mediating conditions from printing through digitization that have rendered each newspaper photograph in digital form. Even in the case of the two images reproduced in the *The Broad Ax*, which were digitized from the very same microfilm reel (reel #00280761059) by the University of Illinois at Urbana-Champaign Library, variations are still apparent. To understand how these subtle differences between images are amplified through digitization, we now turn to optical character recognition. 28

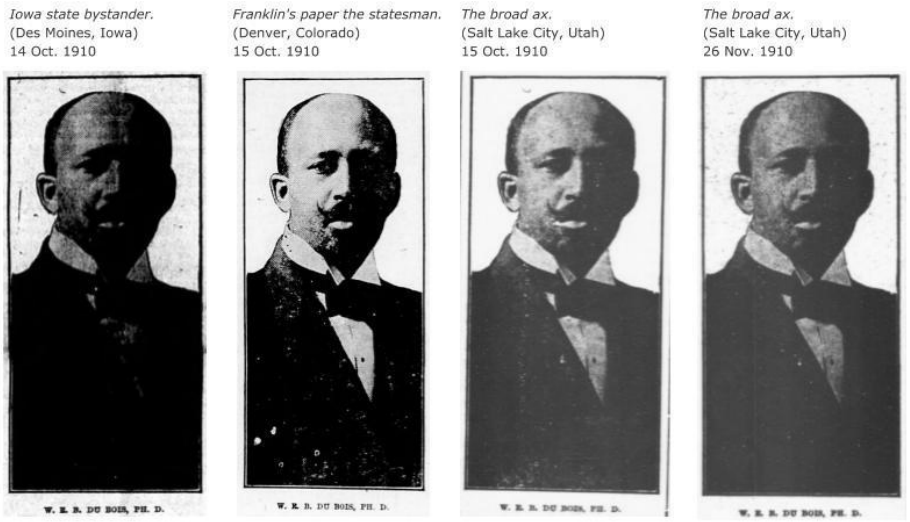

Figure 2. The same image of W.E.B. Du Bois reproduced in 4 different digitized Black newspapers in *Chronicling America* from 1910. Note that the combined effects of printing, microfilming, and digitizing have led to different visual effects in each image, ranging from contrast to sharpness.

V. OCR

Optical character recognition, commonly called OCR, refers to machine learning algorithms that are trained to read images of typewritten text and output machine-readable text, thereby providing the bridge between an image of typewritten text and the transcribed text itself. Because OCR algorithms are "trained and evaluated using labeled data: examples with ground-truth classification labels that have been assigned by another means," the algorithms are considered a form of *supervised learning* in the machine learning literature [Lee 2019]. OCR engines are remarkably powerful in their ability to improve access to historic texts. Indeed, OCR is a crucial form of metadata for *Chronicling America*, enabling keyword search in the search portal and making possible scholarship with the newspaper text at large scales.[12] However, OCR is not perfect. Although humans are able to discern an "E" from an "R" on a digitized page even if the type has been smudged, an OCR engine is not always able to do so: its performance is dependent on factors ranging from the sharpness of text in an image to printing imperfections to the specific typography on the page. 29

In Figure 3, I show the same four images shown in Figure 2, along with OCR transcriptions of the captions provided by *Chronicling America*. All four transcriptions fail to reproduce the true caption with 100% accuracy, differing from one another by at least one character. Consequently, a keyword search of "W. E. B. Du Bois" over the raw text would not register the caption for any of the four photographs (the *Chronicling America* search portal utilizes a form of relevance search to alleviate this problem). These examples reveal how sensitive OCR engines are to slight perturbations, or "noise," in the digitized images, from ink smudges to text sharpness to page contrast. Though the NDNP awardees who contributed these pages may have utilized different OCR engines or chosen different OCR settings, the OCR for the two image captions from *The Broad Ax* that have been digitized from the very same microfilm reel was in all likelihood generated using the same OCR engine and settings. Put succinctly, OCR engines amplify the noise from both the material page and the digitization pipeline.[13] 30

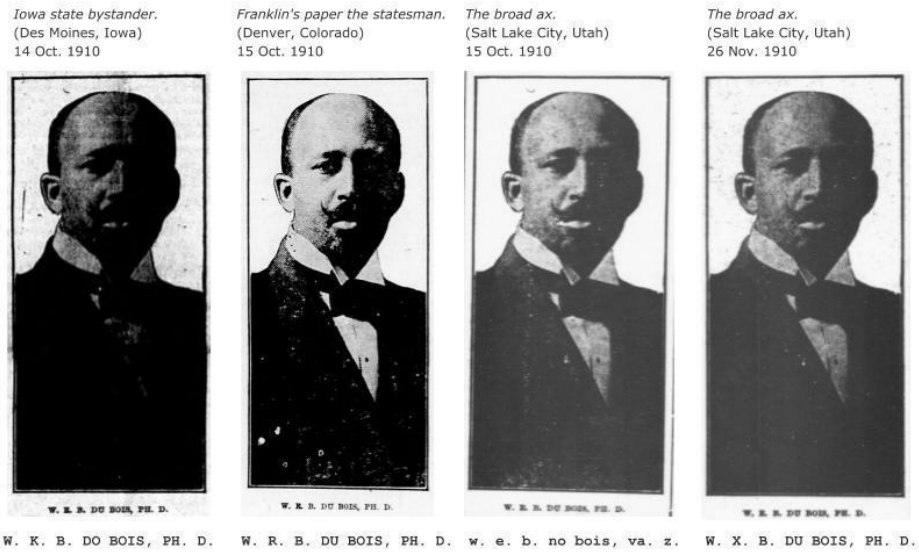

Figure 3. The OCR transcriptions of the caption "W. E. B. DU BOIS, PH. D." appearing in the image of W.E.B. Du Bois reproduced in 4 different digitized Black newspapers in *Chronicling America*. These OCR transcriptions are provided by *Chronicling America*.

Though OCR engines have become standard components of digitization pipelines, it is important to remember that OCR engines are themselves machine learning models that have been trained on sets of transcribed typewritten pages. Like any machine learning model, OCR predictions are thus subject to biases encoded not only in the OCR engine's architecture but also in the training data itself. Though it is often called *algorithmic bias*, this bias is undeniably human, in that the construction of training data machine learning models are imprinted with countless human decisions and judgment calls. For example, if an OCR engine is trained on transcriptions that consistently misspell a word, the OCR engine will amplify this misspelling across all transcriptions of processed pages.[14] A recurring theme of algorithmic bias is that it is a force for marginalization, especially in the context of how we navigate information digitally. In *Algorithms of Oppression*, Safiya Noble describes how Google's search engine consistently marginalizes women and people of color by displaying search results that reinforce racism [Noble 2018]. This bias is not restricted to Google: in *Masked by Trust: Bias in Library Discovery*, Matthew Reidsma articulates how library search engines suffer from similar biases [Reidsma 2019]. Despite the fact that knowledge of algorithmic bias in relation to search engines and image recognition tools is becoming increasingly widespread among the cultural heritage community, the errors introduced by OCR engines are often accepted as inevitable without critical inquiry from this perspective. However, algorithmic bias is a useful framework for examining OCR engines [Alpert-Adams 2016]. 31

Perhaps the most significant challenge to studying OCR engines is that the best-performing and most widely-used OCR engines are proprietary. Though ABBYY FineReader and Google Cloud Vision API offer high performance, the systems fundamentally are black boxes: we have no access to the underlying algorithms or the training data. The ability to audit a system is crucial to developing an understanding of how it works and the biases it encodes. The fact that many OCR engines are opaque prevents us from disentangling whether poor performance on a particular page is due to algorithmic limitations or due to a lack of relevant training data. The distinction is significant: the former may reflect an algorithmic upper bound, whereas the latter reflects decisions made by humans. 32

Indeed, algorithmic bias distorts and occludes the historical record, as it is made discoverable through OCR. Discrepancies in OCR performance for different languages and scripts is a consequence of human prioritization, from the collection of training data and lexicons to the development of the algorithms themselves. As articulated by Hannah Alpert-Abrams in "Machine Reading the *Primeros Libros*," "the machine-recognition of printed characters is a historically charged event, in which the system and its data conspire to embed cultural biases in the output, or to affix them as supplementary information hidden behind the screen" [Alpert-Adams 2016]. Alpert-Abrams's work reveals how the OCR 33

inaccuracies for indigenous languages recorded in colonial scripts perpetuate colonialism. For other languages such as Ladino, typically typeset in Rashi script, the lack of high-performing OCR has presented consistent challenges for digitization and scholarship.

In the case of *Chronicling America*, the National Digital Newspaper Program is exemplary in its efforts to support OCR for non-English languages. In the Notice of Funding Opportunity for the National Digital Newspaper Program produced by the Division of Preservation of Access at the National Endowment for the Humanities, OCR performance in different languages is explicitly addressed: “Applicants proposing to digitize titles in languages other than English must include staff with the relevant language expertise to review the quality of the converted content and related metadata” [NEH Division of Preservation and Access 2020]. I have included this discussion of OCR and algorithmic bias to offer a broader provocation regarding machine learning and digitization: how much text in digitized sources has been transmuted by this effect and thus effectively erased due to inaccessibility when using search and discovery platforms? 34

VI. The Visual Content Recognition Model

I will now turn to the *Newspaper Navigator* pipeline itself, in particular the visual content recognition model. Trained on annotations from the *Beyond Words* crowdsourcing initiative, as well as additional annotations of headlines and advertisements, the visual content recognition model detects photographs, illustrations, maps, comics, editorial cartoons, headlines, and advertisements on historic newspaper pages. 35

As described in the previous section, examining training data is an essential component of auditing any machine learning model, from understanding how the dataset was constructed to uncovering any biases in the composition of the dataset itself. For the visual content recognition model, this examination begins with *Beyond Words*. Launched in 2017 by LC Labs, *Beyond Words* has collected to-date over 10,000 verified annotations of visual content in World War 1-era newspaper pages from *Chronicling America*. The *Beyond Words* workflow consists of the three steps listed below: 36

1. A “Mark” step, in which volunteers are asked to draw bounding boxes around visual content on the page [LC Labs 2017a]. The instructions read as follows:

   > In the Mark step, your task is to identify and select pictures in newspaper pages. For our project, “pictures” means illustrations, photographs, comics, and cartoons. You'll use the marking tool to draw a box around the picture using your mouse. After you have marked all pictures on the newspaper page, click the ‘DONE’ button. Skip the page altogether by clicking the “Skip this page” button. If no illustrations, photographs, or cartoons appear on the page, click the “DONE” button. Not sure if a picture should be marked? Select the “Done for now, more left to mark” button so another volunteer can help finish that page. Please do not select pictures within advertisements.

2. A “Transcribe” step, in which volunteers are asked to transcribe the caption of the highlighted visual content, as well as note the artist and visual content category (“Photograph,” “Illustration,” “Map,” “Comics/Cartoon,” “Editorial Cartoon”) [LC Labs 2017b]. The transcription is pre-populated with the OCR falling within the bounding box in question. The instructions for this step state:

   > Most pictures have captions or descriptions. Enter the text exactly as you see it. Include capitalization and punctuation, but remove hyphenation that breaks words at the end of the line. Use new lines to separate different parts of captions and descriptions. You can zoom in for better looks at the page. You can also select “View the original page” in the upper right corner of the screen to view the original high resolution image of the newspaper.

   An example of this step can be seen in Figure 4.

3. A “Verify” step, in which volunteers are asked to select the best caption for an identified region of visual content from at least two examples; alternatively, a volunteer can add another caption [LC Labs 2017c].

The instructions state:

> Choose the transcription that most accurately captures the text as written. If multiple transcriptions appear valid, choose the first one. If the selected region isn't appropriate for the prompt, click “Bad region”.

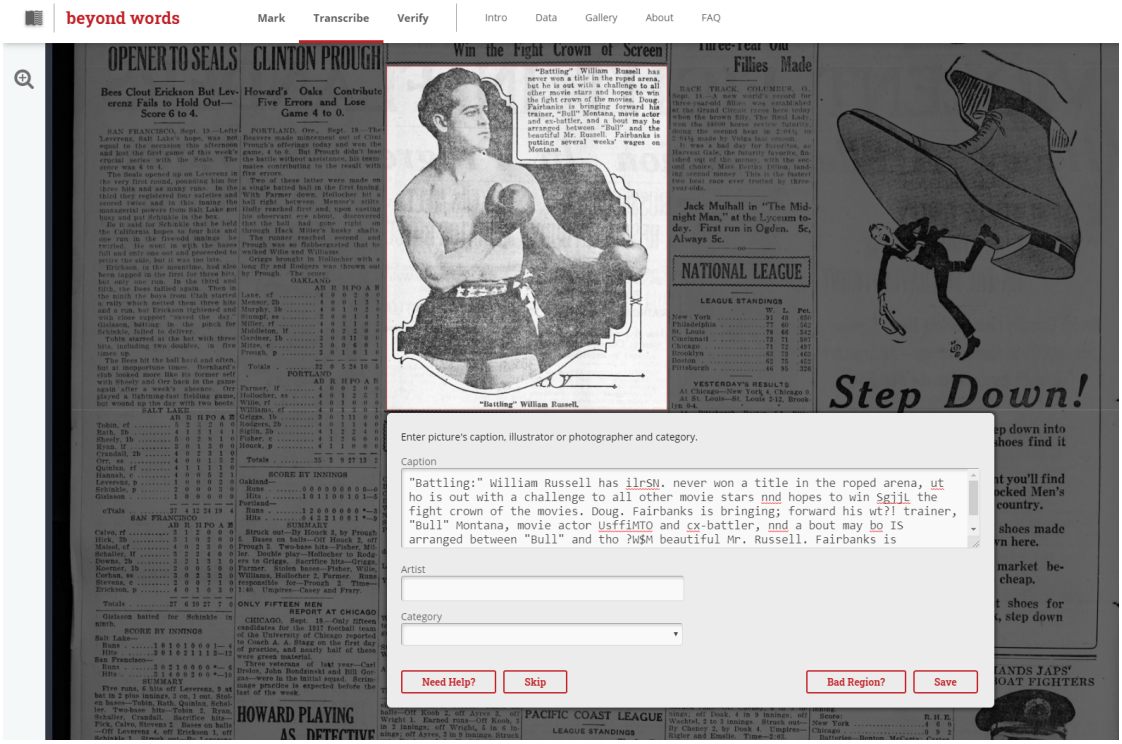

Figure 4. A screenshot showing an example of the “Transcribe” step of the *Beyond Words* workflow. Note that the photograph caption is pre-populated using the OCR falling within the bounding box [LC Labs 2017b].

For the purposes of *Newspaper Navigator*, only the bounding boxes from the “Mark” step and the category labels from the “Transcribe” step were utilized as training data; however, understanding the full workflow is essential because annotations are considered “verified” only if they have passed through the full workflow. 44

A number of factors contribute to which *Chronicling America* pages were processed by volunteers in *Beyond Words*. First, the temporal restriction to World War 1-era pages affects the ability of the visual content recognition model to generalize: after all, if the model is trained on World War 1-era pages, how well should we expect it to perform on 19th century pages? I will return to this question later in the section. Moreover, *Beyond Words* volunteers could select either an entirely random page or a random page from a specific state, an important affordance from an engagement perspective, as volunteers could explore the local histories of states in which they are interested. But this affordance is also imprinted on the training data, as certain states - and thus, certain newspapers - appear at a higher frequency than if the World War-1 era *Chronicling America* pages had been drawn randomly from this temporal range in *Chronicling America*. 45

Furthermore, it should be noted that the “Mark” and “Transcribe” steps - specifically, drawing bounding boxes and labeling the visual content category - are complex tasks. Because newspaper pages are remarkably heterogenous, ambiguities and edge-cases abound. Should a photo collage be marked as one unit or segmented into constituent parts? What precisely is the distinction between an editorial cartoon and an illustration? How much relevant textual content should be included in a bounding box? Naturally, volunteers did not always agree on these choices. In this regard, the notion of a ground-truth, a set of perfect annotations against which we can assess performance, is itself called into question. Moreover, with thousands of annotations, mistakes in the form of missed visual content, as well as misclassifications, are inevitable.[15] These ambiguities and errors are natural components of *any* training dataset and 46

must be taken into account when analyzing a machine learning model's predictions.

A breakdown of *Beyond Words* annotations included in the training data can be found in the second column of Table 1. I downloaded these 6,732 publicly-accessible annotations as a JSON file on December 1, 2019. Table 1 reveals an imbalance between the number of examples for each category; in the language of machine learning, this is called *class imbalance*. While the discrepancy between maps and photographs is to be expected, the fact that so few maps were included was concerning from a machine learning standpoint: a machine learning algorithm's ability to generalize to new data is dependent on having many diverse training examples. To address this concern, I searched *Chronicling America* and identified 134 pages published between January 1st, 1914, and December 31st, 1918, that contain maps. I then annotated these pages myself. 47

In addition, during the development of the *Newspaper Navigator* pipeline, I realized the value in training the visual content recognition model to identify headlines and advertisements. Consequently, I added annotations of headlines and advertisements for all 3,559 pages included in the training data. The statistics for this augmented set of annotations can be found in the third column of Table 1. Though I attempted to use a consistent approach to annotating the headlines and advertisements, my interpretation of what constitutes a headline is certainly not unimpeachable: I am not a trained scholar of periodicals or of print culture; even if I were, the task itself is inevitably subjective. Furthermore, I made decisions to annotate large grids of classified ads as a single ad to expedite the annotation process. Whether this was a correct judgment call can be debated. Lastly, annotating all 3,559 pages for headlines and advertisements required a significant amount of time, and there are inevitably mistakes and inconsistencies embedded within the annotations. My own decisions in terms of how to annotate, as well as my mistakes and inconsistencies, are embedded within the visual content recognition model through training. For those interested in examining the training data directly, the data can be found in the GitHub repository for this project [Lee 2020]. 48

Category	Beyond Words Annotations	Total Annotations
Photograph	4,193	4,254
Illustration	1,028	1,048
Map	79	215
Comic/Cartoon	1,139	1,150
Editorial Cartoon	293	293
Headline	-	27,868
Advertisement	-	13,581
Total	6,732	48,409

Table 1. A breakdown of *Beyond Words* annotations included in the training data for the visual content recognition model, as well as all annotations constituting the training data.

Beyond the construction of the training data, I made manifold decisions regarding the selection of the correct model architecture and the training of the model. Because this discussion surrounding these choices is quite technical, I refer the reader to [Lee et al. 2020] for an in-depth examination. However, I will state that the choice of model, the number of iterations for which the model was trained, and the choice of model parameters are all of significant import for the resulting trained model and consequently, the *Newspaper Navigator* dataset. 49

I will now turn to the visual content recognition model's outputs in relation to the *Newspaper Navigator* pipeline. The model itself consumes a lower-resolution version of a *Chronicling America* page as input and then outputs a JSON file containing predictions, each of which consists of bounding box coordinates,[16] the predicted class (i.e., "photograph", "map", etc.), and a confidence score generated by the machine learning model.[17] Cropping out and saving the visual content required extra code to be written. Because the high-resolution images of the *Chronicling America* pages, in addition to the METS/ALTO OCR, amount to many tens of terabytes of data, questions of data storage became major considerations in the pipeline. I chose to save the extracted visual content as lower-resolution JPEG images in order to 50

reduce the upload time and lessen the storage burden. Though the *Newspaper Navigator* dataset retains identifiers to all high-resolution pages in *Chronicling America,* the images in the *Newspaper Navigator* dataset are altered by the downsampling procedure. This downsampling procedure should be free of any significant biasing effects.

For visual content recognition, “Newspaper Navigator” utilized an object detection model, which is a type of widely-used computer vision technique for identifying objects in images. The performance for computer vision techniques is regularly measured using metrics such as average precision. For “Newspaper Navigator”, the model’s performance on a specific page, as measured by average precision, is dependent on a confluence of factors. These factors include the page’s layout, artifacts and distortions introduced in the microfilming and digitization process, and - most importantly - the composition of the training data. Thus, each image is “seen” differently by the visual content recognition model. In Figure 5, I show the four images of W.E.B. Du Bois, as identified by the visual content recognition model and saved in the *Newspaper Navigator* dataset. Each image is cropped slightly differently. In the case of the image from the *Iowa State Bystander*, extra text is included, while in the case of the images from *The Broad Ax*, the captions are partially cut off. The loss in image quality is due to the aforementioned downsampling performed by the pipeline. This downsampling leads to artifacts such as the dots appearing on Du Bois’s face in the image from the *Iowa State Bystander*, as well as the streaks in the image from *Franklin’s Paper the Statesman*, that are not present in Figure 2. 51

Returning to the question of the visual content recognition model’s performance on pages published outside of the temporal range of the training data (1914-1918), it is possible to provide a quantitative answer by measuring average precision on test sets of annotated pages from different periods of time. In [Lee et al. 2020], I describe this analysis in detail and demonstrate that the performance declines for pages published between 1875 and 1900 and further declines for pages published between 1850 and 1875. This confirms that the composition of the training data directly manifests in the model’s performance. While it is certainly the case that the *Newspaper Navigator* dataset can still be used for scholarship related to 19th century newspapers in *Chronicling America*, any scholarship with the 19th century visual content in the *Newspaper Navigator* dataset must consider how the dataset may skew what visual content is represented. 52

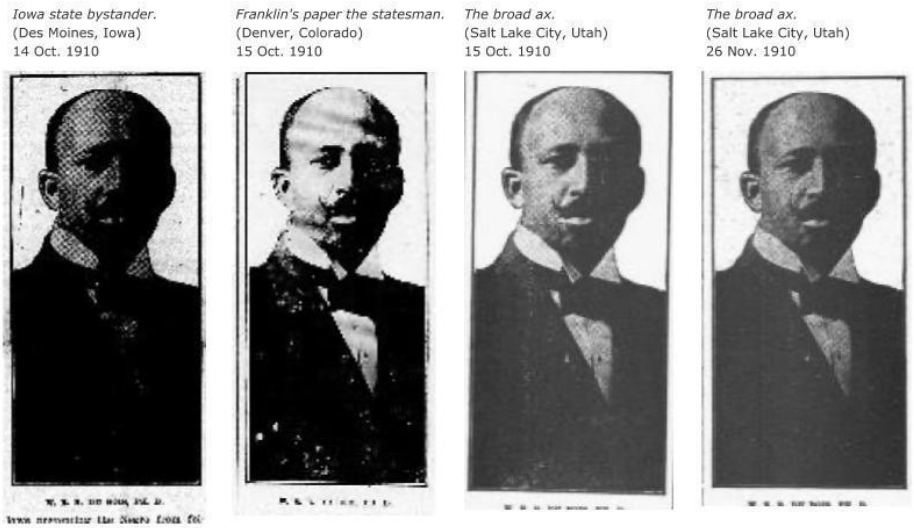

Figure 5. The four images of W.E.B. Du Bois, as identified by the visual content recognition model and included in the *Newspaper Navigator* dataset [Newspaper Navigator 1910a]; [Newspaper Navigator 1910c]; [Newspaper Navigator 1910e]; [Newspaper Navigator 1910g].

Let me conclude this section with a discussion of the act of visual content extraction itself in relation to digitization. While this extraction enables a wide range of affordances for searching *Chronicling America*, it is also an act of decontextualization: visual content no longer appears in relation to the *mise-en-page*. In the Appendix, the full pages 53

containing the photographs of W.E.B. Du Bois are reproduced, showing each photograph in context. Only by examining the full pages does it become clear that the article featuring W.E.B. Du Bois was printed with a second article in the *Iowa State Bystander* and *The Broad Ax*, the headline of which reads: “ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON — Afro-American Women Unite For Active Campaign Against Injustice.” Furthermore, upon examination, the *Iowa State Bystander* front page features the article on *The Crisis* and W.E.B. Du Bois as the most prominent article of the issue. Though links between the extracted visual content and the original *Chronicling America* pages are always retained, this decontextualization inevitably transmutes *how* we perceive and interact with the visual content in *Chronicling America*. Indeed, all uses of machine learning for metadata enhancement are a form of decontextualization, centering the user’s discovery and analysis of content around the metadata itself.

VII. Prediction Uncertainty

Perhaps the most fundamental question to ask of the *Newspaper Navigator* dataset is: “How many photographs does the dataset contain?” Because the dataset has been constructed using a machine learning model, predictions are ultimately probabilistic in nature, quantified by the confidence score returned by the model. This begs the question of what counts as an identified unit of visual content: a user is much more inclined to tally a prediction of a map if it has an associated confidence score of 99% rather than 1%. However, choosing this cut is fundamentally a subjective decision, informed by the user’s end goals with the dataset. In the language of machine learning, picking a stringent confidence cut (i.e., only counting predictions with high confidence scores) emphasizes *precision*: a prediction of a photograph likely corresponds to a true photograph, but the predictions will suffer from false negatives. Conversely, picking a loose confidence cut (i.e., counting predictions with low confidence scores) emphasizes *recall*: most true photographs are identified as such, but the predictions will suffer from many false positives. In this regard, the total number of images in the *Newspaper Navigator* dataset is dependent on one’s desired tradeoff between precision and recall. In Table 2, I show the dynamic range of the dataset size, as induced by three different cuts on confidence score: 90%, 70%, and 50%. Figure 6 shows the effects of different cuts on confidence score for the page featuring W.E.B. Du Bois in the November 26,1910, issue of *The Broad Ax*.

Category	**≥ 90%**	**≥ 70%**	**≥ 50%**
Photograph	1.59×10^6	2.63×10^6	3.29×10^6
Illustration	8.15×10^5	2.52×10^6	4.36×10^6
Map	2.07×10^5	4.59×10^5	7.54×10^5
Comic/Cartoon	5.35×10^5	1.23×10^6	2.06×10^6
Editorial Cartoon	2.09×10^5	6.67×10^5	1.27×10^6
Headline	3.44×10^7	5.37×10^7	6.95×10^7
Advertisement	6.42×10^7	9.48×10^7	1.17×10^8
Total	1.02×10^8	1.56×10^8	1.98×10^8

Table 2. The number of occurrences of each category of visual content in the *Newspaper Navigator* dataset with confidence scores above the listed thresholds (0.9, 0.7, 0.5).

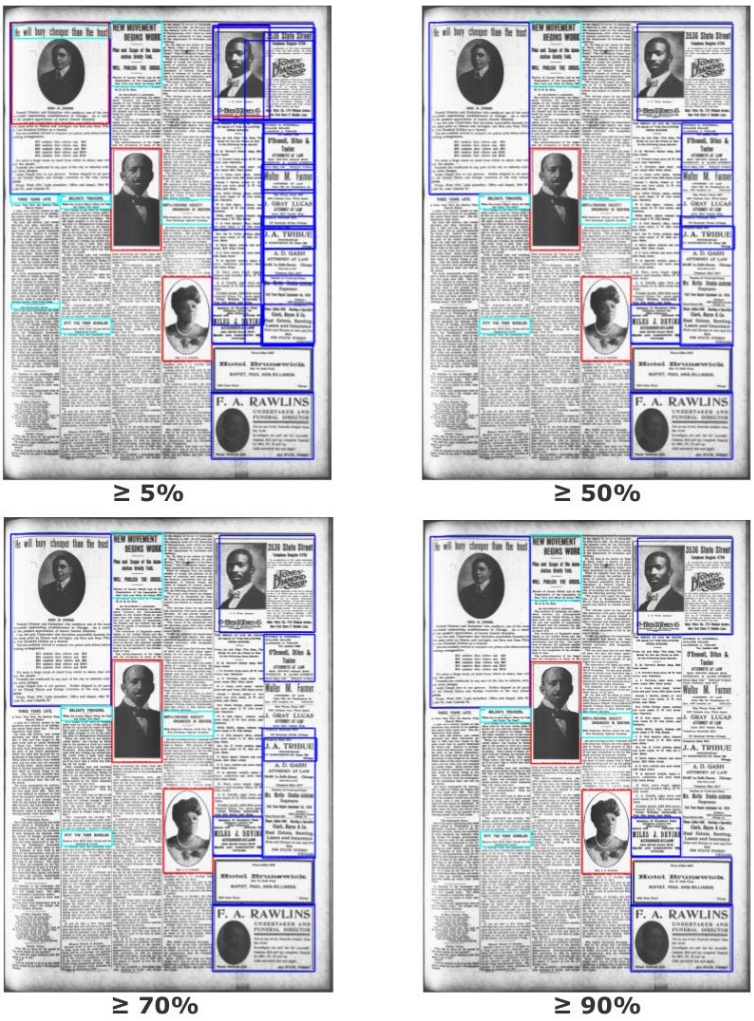

Figure 6. The same page of *The Broad Ax* from November 26, 1910, along with predictions from the visual content recognition model, thresholded on confidence score at 5%, 50%, 70%, and 90% [Newspaper Navigator 1910g]; [Newspaper Navigator 1910h]. Note that red corresponds to a prediction of "photograph", cyan corresponds to a prediction of "headline", and blue corresponds to a prediction of "advertisement".

Rather than pre-selecting a confidence score threshold, the *Newspaper Navigator* dataset contains all predictions with confidence scores greater than 5%,[18] allowing the user to define their own confidence cut when querying the dataset. However, the website for the *Newspaper Navigator* dataset also includes hundreds of pre-packaged datasets in order to make it easier for users to work with the dataset. In particular, users can download zip files containing all of the visual content of a specific type with confidence scores greater than or equal to 90%, for any year from 1850 to 1963. I made this choice of 90% as the threshold cut for these pre-packaged datasets based on heuristic evidence from inspecting sample pre-packaged datasets by eye. However, as articulated above, based on different use cases, this cut of 90% may be too restrictive or permissive: relevant visual content may be absent from the pre-packaged dataset or lost in a sea of other examples. In Figure 7, I show the visual content recognition model's confidence scores for the four images of W.E.B. Du Bois described throughout this data archaeology. The effect of a cut on confidence score can be seen here: selecting a cut of 95% would exclude the image from Franklin's Paper the Statesman. I raise this point to emphasize that even this seemingly innocuous choice of 90% for the pre-packaged datasets alters the discovery process and thus can have an impact on scholarship.

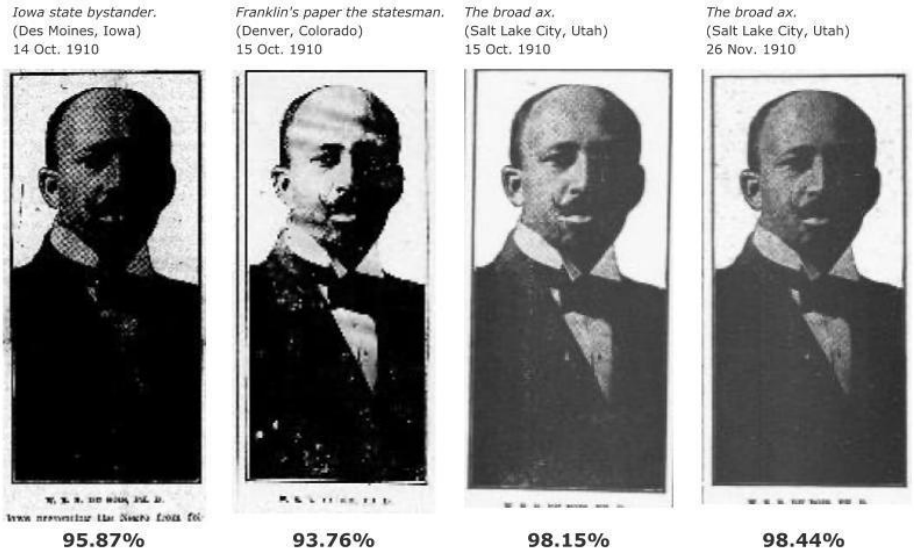

Figure 7. The visual content recognition model's confidence score for each of the four images of W.E.B. Du Bois. Note how the model assigns a different confidence score to each identified image [Newspaper Navigator 1910b]; [Newspaper Navigator 1910d]; [Newspaper Navigator 1910f]; [Newspaper Navigator 1910h].

Just as the bounding box predictions themselves are affected by the training data, as well as newspaper page layout, date of publication, and noise from the digitization pipeline, so too are the confidence scores. In particular, the visual content recognition model suffers from high-confidence misclassifications, for example, crossword puzzles that are identified as maps with confidence scores greater than 90%. High-confidence misclassifications pose challenges for machine learning writ large, and the field of explainable artificial intelligence is largely devoted to developing tools for understanding this type of misclassification [Weld and Bansal 2019]. However, these high-confidence misclassifications can often be traced back to the composition of the training set. For example, the fact that the visual content recognition model sometimes identifies crossword puzzles as maps with high confidence is likely due to the fact that the training data did not contain enough labeled examples of maps and crossword puzzles for the visual content recognition model to differentiate them with high accuracy. 56

The questions surrounding confidence scores and probabilistic descriptions of items is by no means restricted to the *Newspaper Navigator* dataset. I echo Thomas Padilla's assertion that "attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measure of certainty" [Padilla 2019]. Machine-generated metadata such as OCR are also fundamentally probabilistic in nature; this fact is not immediately apparent to end users of cultural heritage collections because cuts on confidence score are typically chosen before surfacing the metadata. Effectively communicating confidence scores, probabilistic descriptions, and the decisions surrounding them to end users remains a challenge for content stewards. 57

VIII. OCR Extraction

In the *Newspaper Navigator* pipeline, a textual description of each prediction is obtained by extracting the OCR within each predicted bounding box. The resulting textual description is thus dependent on not only the OCR provided by *Chronicling America* but also the exact coordinates of the bounding box: if the coordinates of a word in the localized OCR extend beyond the bounds of the box, the word is excluded. I experimented with utilizing tolerance limits to allow words that extend just beyond the bounds of the boxes to be included, but doing so ultimately introduces false positives as well, as words from neighboring articles or visual content were inevitably included some fraction of the time. Once again, the tradeoff between false positives and false negatives is manifest. 58

In Figure 8, I show the textual descriptions of the four images of W.E.B. Du Bois, as identified by the *Newspaper* 59

Navigator pipeline. Significantly, in the *Newspaper Navigator* dataset, the OCR is stored as a list of words, with line breaks removed; these lists are what appear in Figure 8. These four examples provide intuition as to how the captions are altered. While the examples from the *Iowa State Bystander* and *Franklin's Paper the Statesman* both have very similar captions as shown in Figure 3, the captions for both of the examples from *The Broad Ax* are unrecognizable. Because the bounding boxes have clipped the caption, none of the characters from the proper OCR captions from Figure 3 are present. Furthermore, the captions contain OCR noise due to the OCR engine attempting to read text from the photographs. Consequently, the mentions of W.E.B. Du Bois are erased from the textual descriptions in the *Newspaper Navigator* dataset. The visual content in the *Newspaper Navigator* dataset is thus decontextualized not only in the sense that the visual content is extracted from the newspaper pages but also in the sense that the OCR extraction method further alters the textual descriptions. While the images from the *Iowa State Bystander* and *Franklin's Paper the Statesman* are still recoverable with fuzzy keyword search, the two images from *The Broad Ax* are impossible to retrieve with *any* form of keyword search, revealing another instance in which employing automated techniques for collections processing affects discoverability.

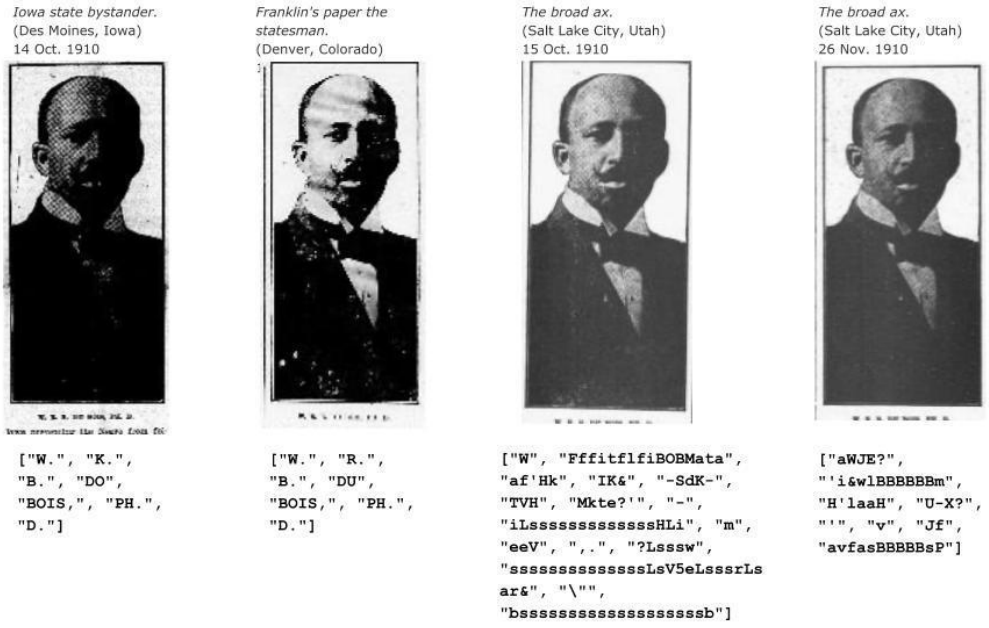

Figure 8. The textual descriptions of each image, as extracted from the OCR and saved in the *Newspaper Navigator* dataset [Newspaper Navigator 1910b]; [Newspaper Navigator 1910d]; [Newspaper Navigator 1910f]; [Newspaper Navigator 1910h].

Fortunately, visual content can still be recovered using similarity search over the images themselves; these methods are discussed in detail in the next section. However, in the case of headlines, the errors introduced by OCR engines and the subsequent OCR extraction have no recourse, as similarity search for images of headlines would only capture similar typography and text layout.[19] 60

To illustrate the effects of this OCR extraction on headlines, I reproduce in Table 3 the extracted OCR as it appears in the *Newspaper Navigator* dataset for Franklin F. Johnson's headline: 61

NEW MOVEMENT 62

BEGINS WORK 63

Plan and Scope of the Asso- 64

ciation Briefly Told. 65

Will Publish the Crisis. 66

Review of Causes Which Led to the

Organization of the Association in

New York and What Its Policy Will

Be-Career and Work of Professor

W.E.B. Du Bois

Iowa State Bystander (14 Oct. 1910)	*Franklin's Paper the Statesman* (15 Oct. 1910)	*The Broad Ax* (15 Oct. 1910)	*The Broad Ax* (26 Nov. 1910)
98.72%	99.57%	99.76%	99.70%
["NEW ", "MOVEMENT ", "BEGINS ", "WORK ", "and ", "Plan ", "Scope ", "of ", "the ", "Asso\u00ad ", "ciation ", "Briefly ", "Told. ", "WILL ", "PUBLISH ", "THE ", "CRISIS. ", "Review ", "of ", "Causae ", "Which ", "Lad ", "to ", "the ", "Organisation ", "of ", "the ", "Auooiation ", "In ", "Naw ", "York ", "and ", "JWhat ", "It* ", "Polioy ", "Will ", "Ba\u2014Career ", "and ", "Wark ", "of ", "Profeasor"]	["NEW ", "MOVEMENT ", "BEGINS ", "WORK ", "Plan ", "and ", "Scope ", "of ", "the ", "Asso ", "ciation ", "Briefly ", "Told. ", "WILL ", "PUBLISH ", "THE ", "CRISIS."]	["NEW ", "MOVEMENT ", "BEGINS ", "WORK ", "Plan ", "and ", "Sep ", "if ", "the ", "Asso ", "ciation ", "Briefly ", "Told. ", "WILL ", "PUBLISH ", "THE ", "CRISIS, ", "Be ", "Career ", "nnd ", "Work ", "of ", "Professor ", "W. ", "E. ", "B. ", "Du ", "Bois. ", "Review ", "of ", "Causes ", "Which ", "Led ", "to ", "the ", "Oraanteallon ", "of ", "th. ", "A.Me!.!?n ", "i ", "i ", "New ", "York ", "and ", "What ", "IU ", "Policy ", "Will"]	["NEW ", "MOVEMENT ", "BEGINS ", "WORK ", "Plan ", "and ", "Scope ", "of ", "the ", "Asso ", "ciation ", "Briefly ", "Told. ", "WILL ", "PUBLISH ", "THE ", "CRISIS. ", "Review ", "of ", "Causes ", "Which ", "Lad ", "to ", "tha ", "Organization ", "of ", "the\" ", "Association ", "In ", "New ", "York ", "and ", "What ", "Its ", "Policy ", "Will"]

Table 3. The extracted OCR associated with each of the four photographs of W.E.B. Du Bois [Newspaper Navigator 1910b]; [Newspaper Navigator 1910d]; [Newspaper Navigator 1910f]; [Newspaper Navigator 1910h].

The full pages are reproduced in the appendix for reference. Notably, all four extracted headlines contain OCR errors, as well as missing words due to the OCR extraction. The visual content recognition model consistently fails to include the last line of the headline, "W.E.B. Du Bois," revealing another case in which Du Bois's name is rendered inaccessible by keyword search in the *Newspaper Navigator* dataset.

IX. Image Embeddings

An *image embedding* canonically refers to a low-dimensional representation of an image, often a list of a few hundred or a few thousand numbers, that captures much of the image's semantic content. Image embeddings are typically generated by feeding an image into a pre-trained neural image classification model (i.e., a model that takes in an image and outputs a label of "dog" or "cat") and extracting a representation of the image from one of the model's hidden layers, often the penultimate layer.[20] Image embeddings are valuable for three reasons:

1. Image embeddings are remarkably adept at capturing semantic similarity between images. For example, images of dogs tend to be clustered together in embedding space, with images of bicycles in another cluster and images of buildings in yet another. These clusters can be fine-grained: sometimes, the red bicycles are grouped closer together than the blue bicycles.
2. Image embeddings can be constructed by feeding images into an image classification model already trained on another dataset (such as ImageNet), meaning that generating image embeddings is a useful method for comparing images without having to construct training data by labeling images.
3. Image embeddings are low-dimensional and thus much smaller in size than the images themselves (i.e., on

the order of kilobytes instead of megabytes). As a result, image embeddings are much less computationally expensive to compare to one another when conducting similarity search, clustering, or related tasks. In short, image embeddings speed up image comparison.

74 Utilizing image embeddings to visualize and explore large collections of images has become an increasingly common approach among cultural heritage practitioners. Projects and institutions that have utilized image embeddings for visualizing cultural heritage collections include the Yale Digital Humanities Lab's PixPlot interface [Yale Digital Humanities Lab 2017], the National Neighbors project [Lincoln et al. 2019], Google Arts and Culture [Google Arts and Culture 2018], The Norwegian National Museum's Principal Components project [Nasjonalmuseet 2017], the State Library of New South Wales's Aero Project [Geraldo 2020], the Royal Photographic Society [Vane 2018], The American Museum of Natural History [Foo 2019], and The National Library of the Netherlands [Lonij and Weavers 2017]; [Weavers and Smits 2020]. These visualizations provide insights into broader themes in the collections, thereby allowing curators, researchers, and the public to explore collections at a scale previously only possible by organizing images by color or other low-level features.[21] In this regard, image embeddings provide new affordances for searching over images that complement canonical faceted and keyword search.

75 Because these image embeddings enable these visualization approaches and open the door to similarity search and recommendation, I opted to include image embeddings as part of the *Newspaper Navigator* pipeline. Indeed, these image embeddings power the similarity search functionality in the *Newspaper Navigator* user interface and, in this regard, are crucial to the broader vision of the project [Lee and Weld 2020].[22] To generate the embeddings, I utilized ResNet-18 and ResNet-50, two variants of a prominent deep learning architecture for image classification, both of which had already been pre-trained on ImageNet [He et al. 2016].

76 ImageNet is perhaps the most well-known image dataset in the history of machine learning. Constructed by scraping publicly available images from the internet and recruiting Amazon Mechanical Turk workers to annotate the images, ImageNet contains approximately 14 million images across 20,000 categories [Deng et al. 2009]; [ImageNet 2020]. Kate Crawford and Trevor Paglen's essay "Excavating AI: The Politics of Images in Machine Learning Training Sets" offers a history and incisive critique of the classification schema of ImageNet; here, I will summarize the most salient critiques. First, many of the categories in the taxonomy utilized are themselves marginalizing [Crawford and Paglen 2019]. Though many of the classes relating to people were removed in 2019, ImageNet had previously bifurcated the "Natural Object > Body > Adult Body" category into "Male Body" and "Female Body" subcategories. Second, ethnic classes were included, implying that 1) classification into rigid categories of ethnicity is possible and appropriate and 2) a machine learning system could learn how to classify ethnicity from these images. Diving deeper, the classifications become horrifying in their supposed granularity: until 2019, an image of a woman in a bikini was accompanied with the tags "slattern, slut, slovenly woman, trollop" [Crawford and Paglen 2019]. Though many embedding models are pre-trained on subsets of ImageNet categories included in the ImageNet Large Scale Visual Recognition Challenge that elide these particularly troubling classifications, these classifications nonetheless necessitate a reckoning with our use of ImageNet writ large, especially in regard to how the semantics of ImageNet are projected onto any image embedding generated with such a model [Russakovsky et al. 2015]. [23]

77 However, questions probing the data in ImageNet fail to critique the ethically questionable practices on which ImageNet is built. Though the researchers responsible for the dataset scraped all 14 million images from public URLs, ImageNet does not provide any guarantees on image copyright, as only the URLs are provided in the database: "The images in their original resolutions may be subject to copyright, so we do not make them publicly available on our server" [ImageNet: What about the Images? 2020]. It is highly unlikely that a photographer with an image in the dataset could have known that a photograph could be used this way, much less actively consent to the image's inclusion, as is the case with subjects in the photographs. Furthermore, the labels themselves were collected using Amazon's Mechanical Turk platform, which has been repeatedly criticized for its exploitative labor practices: as of 2017, workers earned a median wage of approximately $2 an hour on the platform [Haro et al. 2018]. Scholars including Natalia Cecire, Bonnie Mak, and Paul Fyfe have highlighted how outsourced marginalized labor underpins digitization efforts, and the reliance on Mechanical Turk for the production of ImageNet further entrenches the digitization and discovery process within a

system of labor exploitation [Cecire 2011]; [Mak 2017]; [Fyfe 2016]. As cultural heritage practitioners and humanities researchers, we must acknowledge these exploitative practices, and we must reckon with how we perpetuate them through the use of ImageNet as a training source for image search and discovery.

78 In offering these critiques, my intention is not to dismiss ImageNet in a wholesale manner. Certainly, the benefits of utilizing ImageNet are manifold, as evidenced by widespread community adoption, as well as new affordances for searching cultural heritage collections enabled by the dataset that are shaping the contours of digital scholarship. In the case of my own scholarship with Newspaper Navigator, I have elected to utilize machine learning models pre-trained on ImageNet precisely for these reasons. I offer these provocations instead to question how we can do better as a community, not only in imagining alternatives but in bringing them to fruition. Classification is an act of interpretive reduction, whether by human or machine, and thus manifests all too often as an act of oppression.[24] And yet, the structure imposed by classification constitutes the very basis for search and discovery systems. The salient question is thus not how we dispense of these systems but rather how we progressively realize a more inclusive vision of these systems, from the labor practices behind their construction to the very classification taxonomies themselves.

79 How, then, do image embeddings derived from ImageNet mediate our interactions with the photographs in *Newspaper Navigator*? Figure 9 shows a visualization of 1,000 photographs from the *Newspaper Navigator* dataset published during the year 1910. This visualization was created using the ResNet-50 image embeddings, as well as a dimensionality reduction algorithm known as T-SNE [Van der Maaten and Hinton 2009]. With T-SNE, a cluster of photographs indicates that the photographs are likely semantically similar, but the size of the cluster and distances from other clusters bear no meaning [Wattenberg, Viégas, and Johnson 2016]. With this in mind, we can examine the clusters. Despite the fact that the high-contrast, grayscale photographs in *Newspaper Navigator* are markedly different, or “out-of-sample,” in comparison to the clear, color images in ImageNet, the clusters nonetheless capture semantic similarity. In Figure 9, we observe the clustering of photographs depicting crowds of people, as well as photographs depicting ships and the sea. This visualization technique with the image embeddings is thus powerful in helping to navigate large collections of photographs by their semantic content.

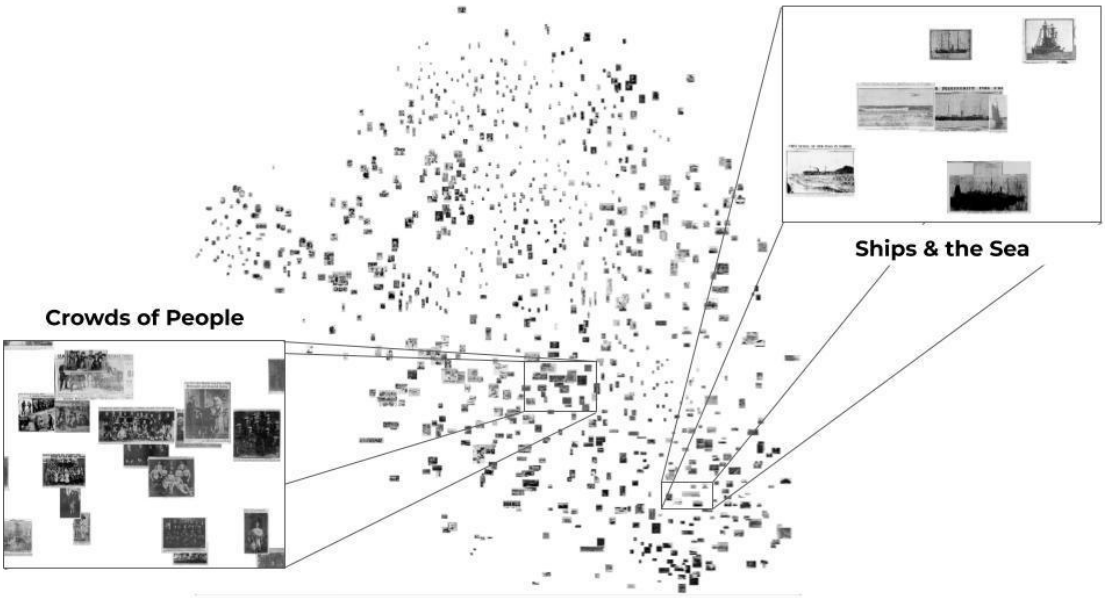

Figure 9. A visualization of 1,000 photographs from the year 1910 in the *Newspaper Navigator* dataset, generated using the *Newspaper Navigator* ResNet-50 image embeddings.

80 What about the photographs of W.E.B. Du Bois? In Figure 10, I show the clusters containing these four photographs. This visualization affords us a lens into the limitations of image embeddings. First, it is evident that image embeddings are directly impacted by the distortions of the digitization process: while the three photographs from *Franklin’s Paper the Statesman* and *The Broad Ax* are clustered together with other portraits, the photograph from the *Iowa State Bystander*

is located in an entirely different cluster - a consequence of the fact that the *Iowa State Bystander* photograph is saturated and that W.E.B. Du Bois's facial features are obscured (notably, neighboring photographs suffer from similar distortions). A search engine powered with these image embeddings would in all likelihood return the three photographs from *Franklin's Paper the Statesman* and *The Broad Ax* together, but the fourth photograph would effectively be lost. This algorithmic mediation is particularly troubling because, as described in Section IV, the microfilming digitization process causes newspaper photographs of darker-skinned people to lose contrast. While this loss in image quality is marginalizing in its own right, image embeddings perpetuate this marginalization: digitized newspaper portraits of darker-skinned individuals are more likely to suffer from saturated facial features, in turn resulting in these photographs being lost during the discovery and retrieval process, as is the case with the saturated *Iowa State Bystander* photograph of W.E.B. Du Bois in Figure 10. Understanding these limitations of image embeddings are particularly relevant in the case of *Newspaper Navigator*, as these image embeddings power the visual similarity search affordance within the publicly-deployed *Newspaper Navigator* search application [Lee and Weld 2020]. Though machine learning methods are often offered as panaceas for automation, this algorithmic erasure reminds us that traditional methods of scholarship and historiography, such as detailed analyses and close readings of Black newspapers in *Chronicling America*, are more important than ever to counter algorithmic bias.

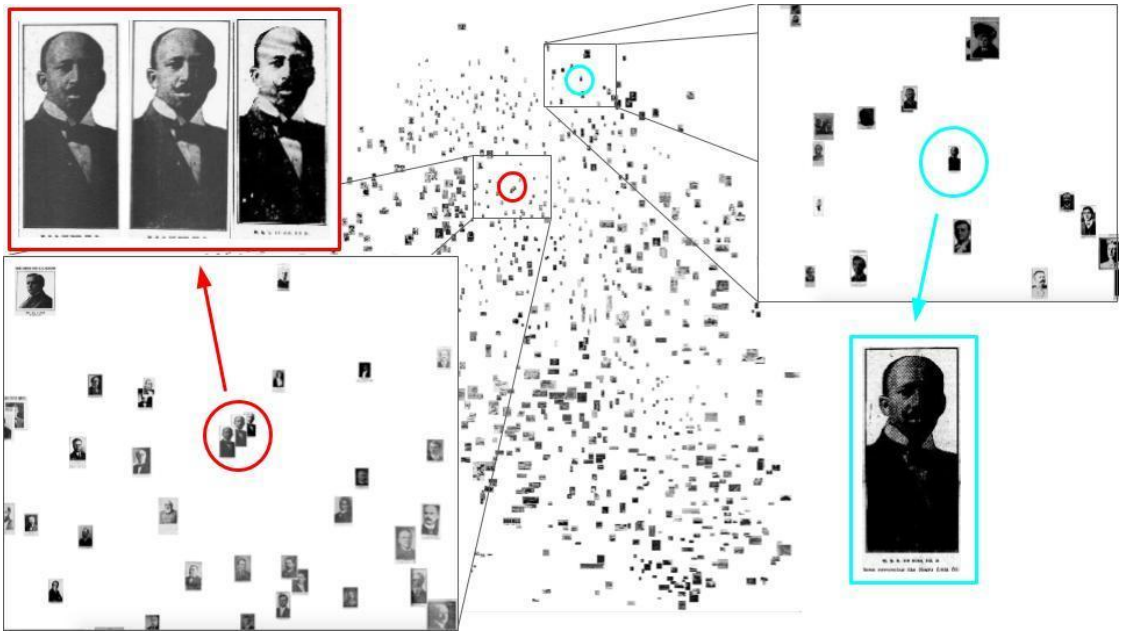

Figure 10. The same visualization as in Figure 9, this time showing the locations of the four photographs of W.E.B. Du Bois.

X. Environmental Impact

Any examination of a dataset whose construction required large-scale computing would be remiss in not investigating the environmental impact of the computation itself. The carbon emissions generated from training a state-of-the-art machine learning model such as BERT is comparable to a single flight across the United States; however, factoring in experimentation and tuning, the carbon emissions can quickly amount to the carbon emissions of a car over its entire lifetime, including fuel [Strubell et al. 2019]. OpenAI's GPT-3 model required several thousand petaflop/s-days to train; without specific numbers, the carbon emissions are not possible to calculate exactly, but they are nonetheless substantial [Brown et al. 2020]. In response, machine learning researchers have recommended ideas such as *Green AI,* with the goal of encouraging the community to value computational efficiency and not just accuracy [Schwartz et al. 2019]. 81

In the case of *Newspaper Navigator*, most of the compute time was devoted to processing all 16.3 million *Chronicling America* pages with the visual content recognition model, as opposed to training the model itself. In Tables 4 and 5, I report details on training the model and running the pipeline, as well as the carbon emissions generated by each step, 82

computed using the Machine Learning Impact Calculator [Lacoste et al. 2019]. In total, approximately 380 kg CO2 were emitted during the construction of the *Newspaper Navigator* dataset, including development, experimentation, training, pipeline processing, and post-processing. It should be noted that this number is an estimate, as the statistics for experimentation and post-processing are difficult to quantify exactly. Nonetheless, this is approximately equivalent to the carbon emissions incurred by a single person flying from Washington, D.C. to Boston [Carbon Footprint Calculator no date]. I include these numbers in the hope that cultural heritage practitioners will consider the environmental impact of utilizing machine learning and artificial intelligence for digital content stewardship. Doing so is essential to the data archaeology: given that climate change will disproportionately affect cultural heritage institutions in regions unable to develop proper infrastructure to withstand rapid temperature fluctuations and unprecedented flooding, even the environmental impacts of utilizing machine learning within digital content stewardship has the capacity to contribute to erasure and marginalization.

Activity	# of NVIDIA T4 GPUs	GPU Hours (each)	Carbon Emissions
Training	1	19	0.96 kg CO2
Pipeline Processing	8	456	144.56 kg CO2
Experimentation for Training and Pipeline Processing (estimate)	8	24	7.66 kg CO2
Total	-	-	153.18 kg CO2

Table 4. Carbon emissions from the GPU usage for *Newspaper Navigator*, broken down by project component. Note that all computation was done on Amazon AWS g4dn instances in the zone "us-east-2". The carbon emissions were calculated using the Machine Learning Impact Calculator [Lacoste et al. 2019].

Activity	CPU Processor (#)	# Processor CPU Cores	CPU Hours (each)	Carbon Emissions
Training	1	4 CPUs	19	1.13 kg CO2
Pipeline Processing	2	48 CPUs	456	181.9 kg CO2
Experimentation for Training and Pipeline Processing (*estimate*)	2	48 CPUs	24	9.57 kg CO2
Extra Computation (dataset post-processing, etc., *estimate*)	1	48 CPUs	168	33.52 kg CO2
Total	-	-	-	226.12 kg CO2

Table 5. Carbon emissions from the CPU usage for *Newspaper Navigator*, broken down by project component. Note that all computation was done on Amazon AWS g4dn instances in the zone "us-east-2". The CPU processors are all 2nd generation Intel Xeon Scalable Processors (Cascade Lake) [Amazon Web Services, Inc. 2020]. The 48-core processor outputs approximately 350 W; the 4-core processor outputs approximately 104 W [Intel 2020a]; [Intel 2020b]. The carbon emissions were calculated using the Machine Learning Impact Calculator [Lacoste et al. 2019]. Note that the energy consumption by RAM is not factored in, but it is insignificant in comparison to the CPU and GPU energy consumption.

XI. Conclusion

In this data archaeology, I have traced four *Chronicling America* pages reproducing the same photograph of W.E.B. Du Bois as they have traveled through the *Chronicling America* and *Newspaper Navigator* pipelines. The excavated genealogy of digital artifacts has revealed the imprintings of the complex interactions between humans and machines. Indeed, the journey of each newspaper page through the *Chronicling America* and *Newspaper Navigator* pipelines is one of refraction, mediation, and decontextualization that is compounded upon with each step. Decisions made decades ago when microfilming a newspaper page inevitably affect how the machine learning models employed for OCR, visual

content extraction, and image embedding generation ultimately process the pages, render them as digital artifacts in the *Newspaper Navigator* dataset, and mediate their discoverability.

As articulated by Trevor Owens in *The Theory and Craft of Digital Preservation*, machine learning and artificial intelligence are the "underlying sciences for digital preservation" [Owens 2018]. Though machine learning techniques provide us with new affordances for searching and studying cultural heritage materials, they have the power to perpetuate and amplify the marginalization and erasure of entire communities within the archive. This erasure, coupled with the labor practices involved in creating training data as well as the environmental impact of training and deploying machine learning models in large-scale digitization pipelines, necessitates that we continue to examine the broader socio-technical ecosystems in which we participate. In doing so, we can work toward a more inclusive vision of the digital collection and the ways in which we render its contents discoverable. 84

How, then, is *Newspaper Navigator* situated within this vision? In reimagining how we search over the visual content in *Chronicling America*, one explicit goal of the project is to engage the public with the rich history preserved within historic American periodicals and thus build on *Chronicling America* as a free-to-use, public domain resource for scholars, educators, students, journalists, genealogists, and beyond [Lee, Berson, and Berson 2021]. [Lee et al. 2021]. With *Newspaper Navigator*, it is my belief that the new modes of interacting with *Chronicling America* have the capacity to not only enable a breadth of new scholarship but also foster engagement in and reckoning with America's multilayered history of oppression. In documenting the different components of the project with this data archaeology and corresponding technical paper [Lee et al. 2020], as well as releasing the full dataset and all code into the public domain, I have intended to be as transparent as possible with the tools and methodologies employed. *Newspaper Navigator* is not without its shortcomings, but my hope is that the project contributes to this vision of the digital collection through transparency and inclusivity, as well as the scholarship and pedagogy that it has enabled. 85

I offer this case study not only to contextualize the *Newspaper Navigator* dataset but also to advocate for the autoethnographic data archaeology as a valuable apparatus for reflecting on a cultural heritage dataset from a humanistic perspective. Though the digital humanities community has yet to adopt the data archaeology as standard practice when creating and releasing cultural heritage datasets, doing so has the capacity to improve accountability and context surrounding applications of machine learning for both practitioners and end users. Given the manifold ways in which machine learning mediates access to the archive and perpetuates erasure, reflecting critically on these systems is not only urgent but essential for transparency and inclusivity. 86

Sources of Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1762114, as well as the Library of Congress Innovator in Residence Position. 87

Acknowledgments

I would like to thank Eileen Jakeway, Jaime Mears, Laurie Allen, Meghan Ferriter, Robin Butterhof, and Nathan Yarasavage at the Library of Congress, as well as Molly Hardy and Joshua Ortiz Baco at the National Endowment for the Humanities, for their thoughtful and enlightening feedback on drafts of this article. I am grateful to my Ph.D. advisor, Daniel Weld, at the University of Washington, for his support, guidance, and invaluable advice with *Newspaper Navigator*. In addition, I would like to thank Kurtis Heimerl and Esther Jang at the University of Washington for the opportunity to formulate and write early sections of this data archaeology as part of this Spring's CSE 599: "Computing for Social Good" course. 88

Lastly, I would like to thank the following people who have shaped *Newspaper Navigator*: Kate Zwaard, Leah Weinryb Grohsgal, Abbey Potter, Chris Adams, Tong Wang, John Foley, Brian Foo, Trevor Owens, Mark Sweeney, and the entire National Digital Newspaper Program staff at the Library of Congress; Devin Naar, Stephen Portillo, Daniel Gordon, and Tim Dettmers at the University of Washington; Michael Haley Goldman, Robert Ehrenreich, Eric Schmalz, and Elliott Wrenn at the United States Holocaust Memorial Museum; Jim Casey at The Pennsylvania State University; Sarah Salter at Texas A&M University-Corpus Christi; and Gabriel Pizzorno at Harvard University. 89

Appendix:

IOWA STATE BYSTANDER.

VOL. XVII, No 18. DES MOINES, IOWA, FRIDAY, OCTOBER 14, 1910. Price Five Cents.

CITY NEWS.

NEW MOVEMENT BEGINS WORK

Plan and Scope of the Association Briefly Told.

WILL PUBLISH THE GRIOT.

ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON.

Restaurants Change Hands.

A New Undertaking Firm.

Getting Together for Victory.

DAVENPORT NOTES.

THE DELINEATOR FOR NOVEMBER.

COUNCIL BLUFFS.

MT. PLEASANT NOTES.

KEOKUK NOTES.

BURLINGTON ITEMS.

BURLINGTON, ILL., NOTES.

CLARINDA.

OTTUMWA NOTES.

CEDAR RAPIDS NOTES.

Figure 11. *Iowa state bystander*. [volume] (Des Moines, Iowa), 14 Oct. 1910. Chronicling America: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn83025186/1910-10-14/ed-1/seq-1/

PAGE 16. THE STATESMAN, DENVER, COLORADO.

From Death Comes Strength

An Injury to One is the Concern of All

INTERNATIONAL LIBERTY UNION

Member Representatives Wanted

NEW MOVEMENT BEGINS WORK

Plan and Scope of the Association Briefly Told.

WILL PUBLISH THE CRISIS.

Review of Causes Which Led to the Organization of the Association in New York and What Its Policy Will Be—Career and Work of Professor W. E. B. Du Bois.

By FRANKLIN S. JOHNSON.

The problem of adjusting the differences between the heterogeneous masses in the United States so that each man will enjoy equality before the law and equality of opportunity for himself and his children has agitated the minds of statesmen oft and on since 1776.

The overthrow of England's sovereignty in the United States and the establishment of a democratic form of government was the first. The existence of slavery, the agitation against it and its final extinction was another step in the recognition of the brotherhood of man.

The enfranchisement of the Negro and the abrogation of many of the laws preventing the Negro from following various fields of activity were steps in recognition of the principles of democracy.

The disfranchisement laws, jim-crow measures, restricted school facilities for the race, are efforts to undo the start made for the extension to all of what many are pleased to term "Jeffersonian principles."

With the curtailment of the rights of the Negro and a steadily growing race consciousness as well as knowledge of public affairs and the growing economic independence of the race has come the conviction that the spirit of democratic principles is no less for the colored man than for the white.

A number of white and colored men, all firm believers in the principles of democracy and strong opposers to all temporizing with social prejudices and its manifestation in discriminating laws, have organized the National Association For the Advancement of Colored People, with offices in the Evening Post building, 20 Vesey street, New York.

The need of such an association can be best gleaned from an address which one of its leading promoters, Oswald Garrison Villard, editor of the New York Evening Post, delivered at the recent meeting in New York of the National Negro Business league. Said Mr. Villard:

"My friends of the Business league, do not let the historian of the future say that in this money making age the colored American became so engrossed in the accumulation of means and the establishment of business as to forget those higher things without which he cannot hope to succeed and rise to great usefulness, to be really worthy of his heritage of American citizenship.

"Do not forget that, while every effort must be spent in fortifying ourselves in every community by business and material success, equal effort must be devoted to that of far greater movement, the insistence upon the Negro's equal and political rights in every place and at all times.

"Anything short of absolute equality before the law is slavery.

"My friends, this republic cannot exist half slave, half free, any more today than it could in the time of Abraham Lincoln."

Dr. W. E. B. Du Bois, the eminent authority on racial questions, has been engaged as secretary of the association. Professor Du Bois is a native of Massachusetts. After graduating from Fisk university in 1888 he entered Harvard university, receiving the degree of bachelor of arts in 1890 and the degree of master of arts the following year.

He was awarded a fellowship and after studying in Berlin, was award

President Young of the Florida Agricultural and Mechanical college is calling special attention of the colored citizens of the state to the fact that it is no longer necessary for them to send their children out of the state for advanced industrial and academic training.

The state is now furnishing them such training tuition free. All thoughtful, taxpaying citizens will avail themselves of this exceptional opportunity

Southern View of Northern Politics.

Politics all over the country is now the topic. Several states will hold important elections. The congressional elections take place in November. All eyes are turned to New York, as will be the outcome of the fight of the progressives and the old guard. The outcome will have much to do with the future of the Republican party.—Savannah (Ga.) Tribune.

to educate their children.

As a matter of fact, the state is now offering th... at the Agricultural and Mechanical college, Tallahassee, free instruction in twenty-three industries, including typewriting and stenography, a collegiate course in science and a course in pedagogy.

T. S. Rector is serving chili piping hot in two ways at his store, 1916 Arapahoe street.

This office is prepared to do all kinds of job printing. Estimates furnished. It is the only race enterprise of its kind operating, and if you desire its service call phone Main 7905. There is a difference between merely soliciting printing and actually doing the work. Get our prices and you will see that difference.

For Rent advertisements appear on page two except such as come to us too late for such positioning. These appear at a cost of 50c per month, or 5c per line if run by the week.

Figure 12. *Franklin's paper the statesman*. (Denver, Colo.), 15 Oct. 1910. Chronicling America: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn91052311/1910-10-15/ed-1/seq-16/

THE BROADAX

NEW MOVEMENT BEGINS WORK

Plan and Scope of the Association Briefly Told.

WILL PUBLISH THE CRISIS

CHIPS

ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON.

Figure 13. *The broad ax.* [volume] (Salt Lake City, Utah), 15 Oct. 1910. Chronicling America: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-10-15/ed-1/seq-2/

Figure 14. *The broad ax*. [volume] (Salt Lake City, Utah), 26 Nov. 1910. Chronicling America: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-11-26/ed-1/seq-3/

Notes

[1] More on the organizational considerations surrounding *Newspaper Navigator* can be found in [Lee et al. 2021].

[2] The public search interface is available at: https://chroniclingamerica.loc.gov/

[3] For more information on the *Beyond Words* workflow, see [LC Labs no date], as well as [Lee et al. 2020].

[4] In particular, the annotations were used to finetune an object detection model that had been pre-trained on Common Objects in Context, a common dataset for benchmarking object detection algorithms.

[5] A screenshot of the workflow can be found later in this article in Figure 4.

[6] For those who are not familiar with image embeddings, a detailed description is provided in Section IX.

[7] For the dataset, see: https://news-navigator.labs.loc.gov; for the code, see https://github.com/LibraryOfCongress/newspaper-navigator.

[8] Indeed, compiling bibliographies of serials published after 1820 remains an immensely difficult task [Hardy and DiCuirci 2019].

[9] The extent to which newspaper microfilming was driven by credible fear of deterioration versus other factors, such as microfilm marketing, is an important question that is rightly debated. For more on this topic, see [Baker 2001].

[10] For example, a 2017 article describing the West Virginia University Libraries' West Virginia & Regional History Center and its participation

in the National Digital Newspaper Program states:“ By August 2017, all known issues of West Virginia’s African-American newspapers from the 19th and early 20th centuries will have been digitized ” [Maxwell 2017]. The article describes Curator Stewart Plein’s efforts to locate surviving copies of three Black West Virginia newspapers in order to digitize and include them in *Chronicling America*.

[11] For a thorough case study of this process, I direct the reader to “Qi-jtb the Raven,” in which Ryan Cordell walks through an example with the Pennsylvania Digital Newspaper Program [Cordell 2017].

[12] For exemplary research collaborations that utilize the *Chronicling America* bulk OCR, see the Viral Text Project and the Oceanic Exchanges Project [Cordell 2017]; [Oceanic Exchanges Project 2017].

[13] For other examinations of how OCR mediates our interactions with digital archives, see [Hitchcock 2013]; [Milligan 2013]; [Strange et al. 2014]; [Traub, van Ossenbruggen, and Hardman 2015]; [Wright 2019].

[14] For a concrete example of a similar phenomenon in the image domain, see [Lee 2019], in which a machine learning algorithm was trained to classify digitized images but consistently misclassified images that had been misoriented 180 degrees in the scanning bed - a consequence of the classifier not having seen enough instances of these misoriented scans during training.

[15] It should be noted that *Beyond Words* was introduced by LC Labs as an experiment, with no interventions in workflow or community management.

[16] Bounding box coordinates refer to the positions of the corners of the predicted bounding box, relative to the image coordinates.

[17] The confidence score is examined in more detail in the next section.

[18] This modest cut is provided to remove the large number of predictions with confidence scores between 0% and 5%, which have high false-positive rates, and thus reduce the size of the *Newspaper Navigator* dataset.

[19] The *Newspaper Navigator* dataset does not retain the cropped images of headlines, as the textual content is more salient than visual snippets in the case of headlines.

[20] If these words are unfamiliar, the three takeaways listed are more important.

[21] For an introduction to some of these methods with lower-level features, see [Manovich 2012].

[22] The search application can be found at: https://news-navigator.labs.loc.gov/search.

[23] The specific categories used in the challenge can be found at: http://image-net.org/challenges/LSVRC/2010/browse-synsets.

[24] For more reading on this topic, see [Bowker and Star 2000].

Works Cited

Alpert-Adams 2016 Alpert-Abrams, H.“Machine Reading the Primeros Libros,”*Digital Humanities Quarterly* 10:4 (2016).

Amazon Web Services, Inc. 2020 “Amazon EC2 Instance Types - Amazon Web Services,” (2020) Amazon Web Services, Inc. Available at: https://aws.amazon.com/ec2/instance-types/. (Accessed: 5 June 2020).

Bailey 2015 Bailey, M. “#transform(Ing)DH Writing and Research: An Autoethnography of Digital Humanities and Feminist Ethics,” *Digital Humanities Quarterly* 9:2 (2015).

Baker 2001 Baker, N. *Double Fold: Libraries and the Assault on Paper*. Random House (2001).

Barrall and Guenther 2005 Barrall, K. and Guenther, C. “Microfilm Selection for Digitization,” (2005). Available at: https://www.loc.gov/ndnp/guidelines/NEH_MicrofilmSelectionNDNP.pdf.

Bender and Friedman 2018 Bender, E., and Friedman, B. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science.” *Transactions of the Association for Computational Linguistics* 6 (2018): 587–604. https://doi.org/10.1162/tacl_a_00041 (Accessed 29 July 2021).

Bowker and Star 2000 Bowker, G., and Star, S. *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge (2000).

Brown et al. 2020 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei D. "Language Models Are Few-Shot Learners,"*ArXiv:2005.14165 [Cs]* (2020), Available at: http://arxiv.org/abs/2005.14165 (Accessed: 6 June 2020).

Carbon Footprint Calculator no date "Carbon Footprint Calculator," Available at: https://calculator.carbonfootprint.com/calculator.aspx?lang=en-GB&tab=3. (Accessed: 6 June 2020).

Cecire 2011 Cecire, N."Works Cited: The Visible Hand," *Works Cited* (blog) (2011). Available at: http://nataliacecire.blogspot.com/2011/05/visible-hand.html.

Chronicling America no date "Chronicling America | Library of Congress," Available at: https://chroniclingamerica.loc.gov/about/ (Accessed 3 July 2020).

Cordell 2017 Cordell, R. "'Q i-Jtb the Raven': Taking Dirty OCR Seriously," *Book History* 20:1, pp. 188–225 (2017). Available at: https://doi.org/10.1353/bh.2017.0006.

Cordell 2020 Cordell, R. "Machine Learning + Libraries: A Report on the State of the Field" (2020). Available at: https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.

Cordell and Smith 2017 Cordell, R., and Smith, D. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (2017), Available at: http://viraltexts.org.

Crawford and Paglen 2019 Crawford, K., and Paglen, T. "Excavating AI: The Politics of Training Sets for Machine Learning" (2019). Available at: https://excavating.ai (Accessed: 19 September 2019).

Deng et al. 2009 Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. "ImageNet: A Large-Scale Hierarchical Image Database,"in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–55, Available at: https://doi.org/10.1109/CVPR.2009.5206848.

Fagan 2016 Fagan, B. "Chronicling White America."American Periodicals: A Journal of History & Criticism 26:1, pp. 10-13 (2016). Available at: https://www.muse.jhu.edu/article/613375.

Farrar 1998 Farrar, H. *The Baltimore Afro-American, 1892-1950*. Greenwood Publishing Group (1998).

Ferriter 2017 Ferriter, M. "Introducing Beyond Words | The Signal," (2017). Available at: //blogs.loc.gov/thesignal/2017/09/introducing-beyond-words/. (Accessed: 13 July 2020).

Foo 2019 Foo, B. "AMNH Photographic Collection," (2020). Available at: https://amnh-sciviz.github.io/image-collection/about.html (Accessed: 11 June 2020).

Franklin's Paper the Statesman 1910 Franklin's paper the statesman. (Denver, Colo.), 15 Oct. 1910. *Chronicling America: Historic American Newspapers*. Library of Congress. Available at: https://chroniclingamerica.loc.gov/lccn/sn91052311/1910-10-15/ed-1/seq-16/

Fyfe 2016 Fyfe, P. "An Archaeology of Victorian Newspapers," *Victorian Periodicals Review* 49:4, pp. 546–77 (2016). Available at: https://doi.org/10.1353/vpr.2016.0039.

Gebru et al. 2020 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. "Datasheets for Datasets."*ArXiv:1803.09010 [Cs]*, March 19, 2020. http://arxiv.org/abs/1803.09010 (Accessed: July 29 2021).

Geraldo 2020 Giraldo, M. "Building Aereo," DX Lab | State Library of NSW (2020). Available at: https://dxlab.sl.nsw.gov.au/blog/building-aereo (Accessed: 2 July 2020).

Google Arts and Culture 2018 "Google Arts & Culture Experiments - t-SNE Map Experiment" (2018). Available at: https://artsexperiments.withgoogle.com/tsnemap/ (Accessed: 11 June 2020).

Hardy and DiCuirci 2019 Hardy, M., and DiCuirci, L. "Critical Cataloging and the Serials Archive: The Digital Making of 'Mill Girls in Nineteenth-Century Print,'" Archive Journal, Available at: http://www.archivejournal.net/?p=8073.

Haro et al. 2018 Hara, K. et al."A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk,"in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (Montreal QC, Canada: Association for Computing Machinery, 2018), pp. 1–14. Available at: https://doi.org/10.1145/3173574.3174023.

He et al. 2016 He, K., Zhang, X., Ren, S., and Sun, J."Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–78, Available at:

https://doi.org/10.1109/CVPR.2016.90.

Hitchcock 2013 Hitchcock, T. “Confronting the Digital,” *Cultural and Social History* 10:1. pp. 9–23 (2013). Available at: https://doi.org/10.2752/147800413X13515292098070.

Holland et al. 2018 Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K.“The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.” *ArXiv:1805.03677 [Cs]*, May 9, 2018. http://arxiv.org/abs/1805.03677 (Accessed 29 July 2021).

ImageNet 2020 “ImageNet, ”Available at: http://image-net.org/index (Accessed: 8 June 2020).

ImageNet: What about the Images? 2020 “What about the images?”Available at: http://image-net.org/download-faq (Accessed: 8 June 2020).

Intel 2020a “Intel® Xeon® Platinum 9242 Processor (71.5M Cache, 2.30 GHz) Product Specifications,” Available at: https://ark.intel.com/content/www/us/en/ark/products/194145/intel-xeon-platinum-9242-processor-71-5m-cache-2-30-ghz.html (Accessed: 5 June 2020).

Intel 2020b “Intel® Xeon® Platinum 8256 Processor (16.5M Cache, 3.80 GHz) Product Specifications,” Available at: https://ark.intel.com/content/www/us/en/ark/products/192467/intel-xeon-platinum-8256-processor-16-5m-cache-3-80-ghz.html (Accessed: June 5, 2020).

Iowa State Bystander 1910 Iowa state bystander. [volume] (Des Moines, Iowa), 14 Oct. 1910. *Chronicling America: Historic American Newspapers*. Library of Congress. Available at: https://chroniclingamerica.loc.gov/lccn/sn83025186/1910-10-14/ed-1/seq-1/

LC Labs 2017a LC Labs, “Beyond Words: Mark”Available at: http://beyondwords.labs.loc.gov/#/mark (Accessed 5 June, 2020).

LC Labs 2017b LC Labs, “Beyond Words: Transcribe,” Available at: http://beyondwords.labs.loc.gov/#/transcribe (Accessed 5 June, 2020).

LC Labs 2017c LC Labs, “Beyond Words: Veriffy,” Available at: http://beyondwords.labs.loc.gov/#/verify (Accessed 5 June, 2020).

LC Labs and Digital Strategy Directorate 2020 LC Labs and Digital Strategy Directorate, “Machine Learning + Libraries Summit Event Summary”(2020). Available at: https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

LC Labs no date LC Labs, Beyond Words | Experiments. Available at: https://labs.loc.gov/work/experiments/beyond-words/ (Accessed 5 June, 2020).

Lacoste et al. 2019 Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. “Quantifying the Carbon Emissions of Machine Learning,” *ArXiv:1910.09700 [Cs]* (2019). Available at: http://arxiv.org/abs/1910.09700.

Lee 2019 Lee, B. “Machine Learning, Template Matching, and the International Tracing Service Digital Archive: Automating the Retrieval of Death Certificate Reference Cards from 40 Million Document Scans,” *Digital Scholarship in the Humanities* 34:3, pp. 513-535 (2019). Available at: https://doi.org/10.1093/llc/fqy063.

Lee 2020 Lee, B. *LibraryOfCongress/Newspaper-Navigator*, GitHub Repository (Library of Congress, 2020). Available at: https://github.com/LibraryOfCongress/newspaper-navigator.

Lee and Weld 2020 Lee, B., and Weld, D. “Newspaper Navigator: Open Faceted Search for 1.5 Million Images,” UIST '20 Adjunct: Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp. 120-122 (2020). Available at: https://doi.org/10.1145/3379350.3416143.

Lee et al. 2020 Lee, B., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., and Weld, D. “The Newspaper Navigator Dataset: Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America,” CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management , pp. 3055–3062 (2020). Available at: https://doi.org/10.1145/3340531.3412767.

Lee et al. 2021 Lee, B., Mears, J., Jakeway, E., Ferriter, M., and Potter, A. “Newspaper Navigator: Putting Machine Learning in the Hands of Library Users,”*EuropeanaTech Insight* 16 (2021). Available at: https://pro.europeana.eu/page/issue-16-newspapers.

Lee, Berson, and Berson 2021 Lee, B., Berson, I., and Berson, M. “Machine Learning and the Social Studies,” *Social Education* 85:2, pp. 88-92 (2021). Available at: https://www.socialstudies.org/social-education/85/2/machine-learning-and-social-studies.

Library of Congress 2019 "Digital Strategy | Library of Congress," Library of Congress (2019). Available at: https://www.loc.gov/digital-strategy/ (Accessed: 30 May 2020).

Lincoln et al. 2019 Lincoln, M., Levin, G., Conell, S., and Huang, L. (2019) "National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections" (2019) Available at: https://nga-neighbors.library.cmu.edu. (Accessed: 30 May 2020).

Lonij and Weavers 2017 Lonij, J., and Wevers, M. (2017) SIAMESE. KB Lab: The Hague (2017). Available at: http://lab.kb.nl/tool/siamese.

Lorang et al 2020 Lorang, E., Soh, L., Liu, Y., and Pack, C. "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project" (2020). Available at: https://digitalcommons.unl.edu/libraryscience/396/.

Mak 2017 Mak, B. "Archaeology of a Digitization," *Journal of the Association for Information Science and Technology* 65:8, pp. 1515–26 (2014). Available at: https://doi.org/10.1002/asi.23061.

Manovich 2012 Manovich, L. "How to Compare One Million Images?," in *Understanding Digital Humanities*, ed. David M. Berry (London: Palgrave Macmillan UK, 2012), pp. 249–78. Available at: https://doi.org/10.1057/9780230371934_14.

Maxwell 2017 Maxwell, M. "WVU Today | WVRHC Seeking Copies of Rare African-American Newspapers" (2017). Available at: https://wvutoday.wvu.edu/stories/2017/01/19/wvrhc-seeking-copies-of-rare-african-american-newspapers. (Accessed 11 July 2020).

Mears 2014 Mears, J. *National Digital Newspaper Program Impact Study 2004-2014*, National Endowment for the Humanities (2014). Available at: https://www.neh.gov/divisions/preservation/featured-project/neh-releases-national-digital-newspaper-program-impact-study. (Accessed 29 May 2020).

Meta | Morphosis no date "Meta | Morphosis: Tutorials," National Digital Newspaper Program and the University of Kentucky Libraries. Available at: https://www.uky.edu/Libraries/NDNP/metamorphosis/tutorials.html (Accessed 3 July 2020).

Milligan 2013 Milligan, I. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review* 94:4, pp. 540–69 (2013). Available at: https://doi.org/10.3138/chr.694.

Mitchell et al. 2019 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I., and Gebru, T. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 29, 2019, 220–29. https://doi.org/10.1145/3287560.3287596.

NEH Division of Preservation and Access 2020 Division of Preservation and Access (NEH), "Notice of Funding Opportunity, National Digital Newspaper Program" (2020). Available at: https://www.neh.gov/sites/default/files/inline-files/National-Digital-Newspaper-Program-NOFO-January-2020.pdf. (Accessed 28 June 2020).

Nasjonalmuseet 2017 "Project: «Principal Components»," Nasjonalmuseet (2018). Available at: https://www.nasjonalmuseet.no/en/about-the-national-museum/collection-management — -behind-the-scenes/digital-collection-management/project-principal-components/ (Accessed 11 June 2020).

National Digital Newspaper Program 2019 "About the Program - National Digital Newspaper Program (Library of Congress)," (2019). Available at: https://www.loc.gov/ndnp/about.html (Accessed 3 July 2020).

National Digital Newspaper Program 2020 The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants 2020-22 Awards (2020). Available at: https://www.loc.gov/ndnp/guidelines/ (Accessed 28 June 2020).

National Digital Newspaper Program no date "Content Selection - National Digital Newspaper Program (Library of Congress)"(2020). Available at: https://www.loc.gov/ndnp/guidelines/selection.html (Accessed 3 July 2020).

Newspaper Navigator 1910a Image of W.E.B. Du Bois from the *Iowa State Bystander* (14 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iahi_ames_ver01/data/sn83025186/00202198417/1910101401/1015/001_0_95.jpg.

Newspaper Navigator 1910b [Newspaper Navigator 1910b] *Newspaper Navigator* metadata for the *Iowa State Bystander* (14 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iahi_ames_ver01/data/sn83025186/00202198417/1910101401/1015.json.

Newspaper Navigator 1910c Image of W.E.B. Du Bois from *Franklin's Paper the Statesman* (15 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-

navigator.labs.loc.gov/data/cohi_abbeyville_ver01/data/sn91052311/00279550730/1910101501/2272/001_0_93.jpg

Newspaper Navigator 1910d *Newspaper Navigator* metadata for *Franklin's Paper the Statesman* (15 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/cohi_abbeyville_ver01/data/sn91052311/00279550730/1910101501/2272 .json

Newspaper Navigator 1910e Image of W.E.B. Du Bois from *The Broad Ax* (15 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910101501/0538/002_0_98.jpg

Newspaper Navigator 1910f *Newspaper Navigator* metadata for *The Broad Ax* (15 October 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910101501/0538.json

Newspaper Navigator 1910g Image of W.E.B. Du Bois from The Broad Ax (26 November 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910112601/0564/004_0_98.jpg

Newspaper Navigator 1910h *Newspaper Navigator* metadata for The Broad Ax (26 November 1910). From the Library of Congress, Newspaper Navigator dataset: Extracted Visual Content from Chronicling America. Available at: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910112601/0564.json

Noble 2018 Noble, S. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, New York (2018).

Oceanic Exchanges Project 2017 Oceanic Exchanges Project Team. Oceanic Exchanges: Tracing Global Information Networks In Historical Newspaper Repositories, 1840-1914 (2017). Available at: 10.17605/OSF.IO/WA94S.

Owens 2018 Owens, T. *The Theory and Craft of Digital Preservation*. Johns Hopkins University Press, Baltimore (2018).

Owens and Padilla 2020 Owens, T., and Padilla, T. "Digital Sources and Digital Archives: Historical Evidence in the Digital Age," *International Journal of Digital Humanities* (2020). Available at:https://doi.org/10.1007/s42803-020-00028-7 https://doi.org/10.1007/s42803-020-00028-7.

Padilla 2019 Padilla, T. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (2019). Available at: https://doi.org/10.25333/xk7z-9g97.

Reidsma 2019 Reidsma, M. *Masked by Trust: Bias in Library Discovery.* Litwin Books, Sacramento (2019).

Reisman et al. 2018 Reisman, D., Schultz, J., Crawford, K., Whittaker, M. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability* (2018). Available at: https://ainowinstitute.org/aiareport2018.pdf.

Russakovsky et al. 2015 Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. "ImageNet Large Scale Visual Recognition Challenge,"*International Journal of Computer Vision* 115:3, pp. 211-252 (2015). Available at:https://doi.org/10.1007/s11263-015-0816-y https://doi.org/10.1007/s11263-015-0816-y.

Schwartz et al. 2019 Schwartz, R., Dodge, J., Smith, N., and Etzioni, O. "Green AI," ArXiv:1907.10597 [Cs, Stat], (2019). Available at:http://arxiv.org/abs/1907.10597 http://arxiv.org/abs/1907.10597.

Strange et al. 2014 Strange, C., McNamara, D., Wodak, J., and Wood, I. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers," *Digital Humanities Quarterly* 8:1 (2014). Available at: http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html.

Strubell et al. 2019 Strubell, E., Ganesh, A., and McCallum, A. "Energy and Policy Considerations for Deep Learning in NLP," ArXiv:1906.02243 [Cs] (2019). Available at: http://arxiv.org/abs/1906.02243.

The Broad Ax 1910a The broad ax. [volume] (Salt Lake City, Utah), 15 Oct. 1910. *Chronicling America: Historic American Newspapers*. Library of Congress. Available at: https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-10-15/ed-1/seq-2/

The Broad Ax 1910b The broad ax. [volume] (Salt Lake City, Utah), 26 Nov. 1910. *Chronicling America: Historic American Newspapers*. Library of Congress. Available at: https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-11-26/ed-1/seq-3/

Traub, van Ossenbruggen, and Hardman 2015 Traub, M., van Ossenbruggen, J., and Hardman, L. "Impact Analysis of OCR Quality on Research Tasks in Digital Archives," in *Research and Advanced Technology for Digital Libraries*, ed. Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla (Cham: Springer International Publishing, 2015), 252–263.

Van der Maaten and Hinton 2009 van der Maaten, L., and Hinton, G. “Visualizing Data Using T-SNE,” *Journal of Machine Learning Research* 9, pp. 2579-2605 (2008). Available at: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Vane 2018 Vane, O. “Visualising the Royal Photographic Society Collection: Part 2 • V&A Blog,” *V&A Blog* (2018). Available at: https://www.vam.ac.uk/blog/digital/visualising-the-royal-photographic-society-collection-part-2.

Wattenberg, Viégas, and Johnson 2016 Wattenberg, M., Viégas, F., and Johnson, I. “How to Use T-SNE Effectively,” *Distill* 1:10 (2016). Available at: https://doi.org/10.23915/distill.00002.

Weavers and Smits 2020 Wevers, M., and Smits, T. “The Visual Digital Turn: Using Neural Networks to Study Historical Images,” *Digital Scholarship in the Humanities* 35:1, pp. 194-207 (2020). Available at: https://doi.org/10.1093/llc/fqy085.

Weld and Bansal 2019 Weld, D., and Bansal, G. 2019. The challenge of crafting intelligible intelligence. Commun. ACM 62: 6, pp. 70–79 (2019). Available at: https://doi.org/10.1145/3282486.

Williams 2019 Williams, L. “What Computational Archival Science Can Learn from Art History and Material Culture Studies,” in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3153–55. Available at: https://doi.org/10.1109/BigData47090.2019.9006527.

Wright 2019 Wright, R. “Typewriting Mass Observation Online: Media Imprints on the Digital Archive,” *History Workshop Journal* 87, pp. 118–38 (2019). Available at: https://doi.org/10.1093/hwj/dbz005.

Yale Digital Humanities Lab 2017 “Yale Digital Humanities Lab - PixPlot” (2020). Available at: https://dhlab.yale.edu/projects/pixplot/ (Accessed 11 June 2020).