The Newspaper Navigator Dataset:
Extracting And Analyzing Visual Content from 16 Million Historic Newspaper Pages in Chronicling America

Benjamin Lee[*]
University of Washington
Library of Congress
bcgl@cs.washington.edu

Jaime Mears
LC Labs
Library of Congress

Eileen Jakeway
LC Labs
Library of Congress

Meghan Ferriter
LC Labs
Library of Congress

Chris Adams
IT Design & Development
Library of Congress

Nathan Yarasavage
National Digital Newspaper
Program
Library of Congress

Deborah Thomas
National Digital Newspaper
Program
Library of Congress

Kate Zwaard
Digital Strategy & LC Labs
Library of Congress

Daniel Weld
University of Washington

ABSTRACT

Chronicling America is a product of the National Digital Newspaper Program, a partnership between the Library of Congress and the National Endowment for the Humanities to digitize historic newspapers. Over 16 million pages of historic American newspapers have been digitized for Chronicling America to date, complete with high-resolution images and machine-readable METS/ALTO OCR. Of considerable interest to Chronicling America users is a semantified corpus, complete with extracted visual content and headlines. To accomplish this, we introduce a visual content recognition model trained on bounding box annotations of photographs, illustrations, maps, comics, and editorial cartoons collected as part of the Library of Congress's Beyond Words crowdsourcing initiative and augmented with additional annotations including those of headlines and advertisements. We describe our pipeline that utilizes this deep learning model to extract 7 classes of visual content: headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements, complete with textual content such as captions derived from the METS/ALTO OCR, as well as image embeddings for fast

image similarity querying. We report the results of running the pipeline on 16.3 million pages from the Chronicling America corpus and describe the resulting Newspaper Navigator dataset, the largest dataset of extracted visual content from historic newspapers ever produced. The Newspaper Navigator dataset, finetuned visual content recognition model, and all source code are placed in the public domain for unrestricted re-use.

INTRODUCTION

Chronicling America, an initiative of the National Digital Newspaper Program - itself a partnership of the Library of Congress and the National Endowment for the Humanities - is an invaluable resource for academic, local, and public historians; educators and students; genealogists; journalists; and members of the public to explore American history through the uniquely rich content preserved within historic local newspapers. Over 16 million pages of newspapers published between 1789 to 1963 are publicly available online through a search portal, as well as via a public API. Among the page-level data are 400 DPI images, as well as METS/ALTO OCR, a standard maintained by the Library of Congress that includes text localization [2].

The 16.3 million Chronicling America pages included in the Newspaper Navigator cover 174 years of American history, inclusive of 47 states, Washington, D.C., and Puerto Rico. In Figure 1, we show choropleth maps displaying the geographic coverage of the 16.3 million Chronicling America newspaper pages included in the Newspaper Navigator dataset. In Figure 2, we show the temporal coverage of these pages. The coverage reflects the selection process for determining which newspapers to include in Chronicling America; for an in-depth

---

[*]Work conducted while an Innovator-in-Residence at the Library of Congress and Ph.D. student in the Paul G. Allen School for Computer Science and Engineering at the University of Washington.

examination, please refer to [26, 55]. The selection process should be considered in the methodology of any research performed using the Newspaper Navigator dataset.

While the images and OCR in Chronicling America provide a wealth of information, users interested in extracted visual content, including headlines, are currently restricted to general keyword searches or manual searches over individual pages in Chronicling America. For example, staff at the Library of Congress have produced a collection of Civil War maps in historic newspapers to date, but the collection is far from complete due to the difficulty of manually searching over the hundreds of thousands of Chronicling America pages from 1861 to 1865 [7]. A complete dataset would be of immense value to historians of the Civil War. Likewise, collecting all of the comic strips from newspapers published in the early 20th century would provide comic researchers with a corpus of unprecedented scale. In addition, users currently have no reliable method of determining what disambiguated articles appear on each page, presenting challenges for natural language processing (NLP) approaches to studying the corpus. A dataset of extracted headlines not only gives researchers insight into the individual articles that appear on each page but also enables users to ask questions such as, "Which news topics appeared above the fold versus below the fold in what newspapers?" Indeed, the digital humanities questions that could be asked with such a dataset abound. And yet, the possibilities extend beyond the digital humanities to include public history, creative computing, educational use within the classroom, and public engagement with the Library of Congress's collections.

To begin the construction of larger datasets of visual content within Chronicling America and to engage the American public, the Library of Congress Labs launched a crowdsourcing initiative called Beyond Words[1] in 2017. With this initiative, volunteers were asked to draw bounding boxes around photographs, illustrations, comics, editorial cartoons, and maps in World War 1-era newspapers in Chronicling America; they were also asked to transcribe captions by correcting the OCR within each bounding box annotation, as well as record the content creator. Approximately 10,000 verified Beyond Words annotations have been collected to date.

Our research builds on the crowdsourced Beyond Words annotations by utilizing the bounding boxes drawn around photographs, illustrations, comics, editorial cartoons, and maps, as well as additional annotations including ones marking headlines and advertisements, to finetune a pre-trained Faster-RCNN implementation from Detectron2's Model Zoo [58, 76]. Our visual content recognition model predicts bounding boxes around these 7 different classes of visual content in historic newspapers. This paper presents our work training this visual content recognition model and constructing a pipeline for automating the identification of this visual content in Chronicling America newspaper pages. Drawing inspiration from the Beyond Words workflow, we extract corresponding textual content such as headlines and captions by identifying text from the METS/ALTO OCR that falls within each predicted bounding box. This method is effective at captioning

because Beyond Words volunteers were asked to include captions and relevant textual content within their bounding box annotations. Lastly, in order to enable fast similarity querying for search and recommendation tasks, we generate image embeddings for the extracted visual content using ResNet-18 and ResNet-50 models pre-trained on ImageNet. This resulting dataset, which we call the Newspaper Navigator dataset, is the largest collection of extracted visual content from historic newspapers ever produced.

Our contributions are as follows:

1. We present a publicly available pipeline for extracting visual and textual content from historic newspaper pages, designed to run at scale over terabytes of image data. Visual content categories include headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements.

2. We release into the public domain a finetuned Faster-RCNN model for this task that achieves 63.4% bounding box mean average precision (mAP)[2] on a validation set of World War 1-era Chronicling America pages.

3. We present the Newswpaper Navigator dataset, a new public dataset of extracted headlines and visual content, as well as corresponding textual content such as titles and captions, produced by running the pipeline over 16.3 million historic newspaper pages in Chronicling America. This corpus represents the largest dataset of its kind ever produced.

RELATED WORK

Corpora & Datasets

Over the past 15 years, efforts across the world to digitize historic newspapers have been remarkably successful [52]. In addition to Chronicling America, examples of large repositories of digitized newspapers include Trove [23], Europeana [54, 75], Delpher [3], The British Newspaper Archive [6], OurDigitalWorld [12], Papers Past [13], NewspaperSG [22], newspapers.com [5] and Google Newspaper Search [24]. The availability of newspapers at the scale of millions of digitized pages has inspired the construction of datasets for supervised learning tasks related to digitized newspapers. In addition to Beyond Words, datasets for historic newspaper recognition include the National Library of Luxembourg's historic newspaper datasets [8] that include segmented articles and advertisements; CHRONIC, a dataset of 452,543 images in historic Dutch newspapers [67]; and Europeana's SIAMESET, a dataset of 426,777 advertisements in historic Dutch newspapers [73]. Datasets for machine learning tasks with historical documents include READ-BAD [33] and DIVA-HisDB [61]. However, all of these datasets are designed to serve as training sets rather than as comprehensive datasets of extracted content from full corpora. Our work instead seeks to use the Beyond Words dataset to train a visual content recognition model in order to process the visual content in the Chronicling America corpus comprising 16+ million historic newspaper pages.

---

[1] https://labs.loc.gov/work/experiments/beyond-words/

[2]Mean average precision is the standard metric used for benchmarking object detection models, incorporating intersection over union to assess precision and recall. We describe the metric in more detail in Section 6.

(a) State-level choropleth map.

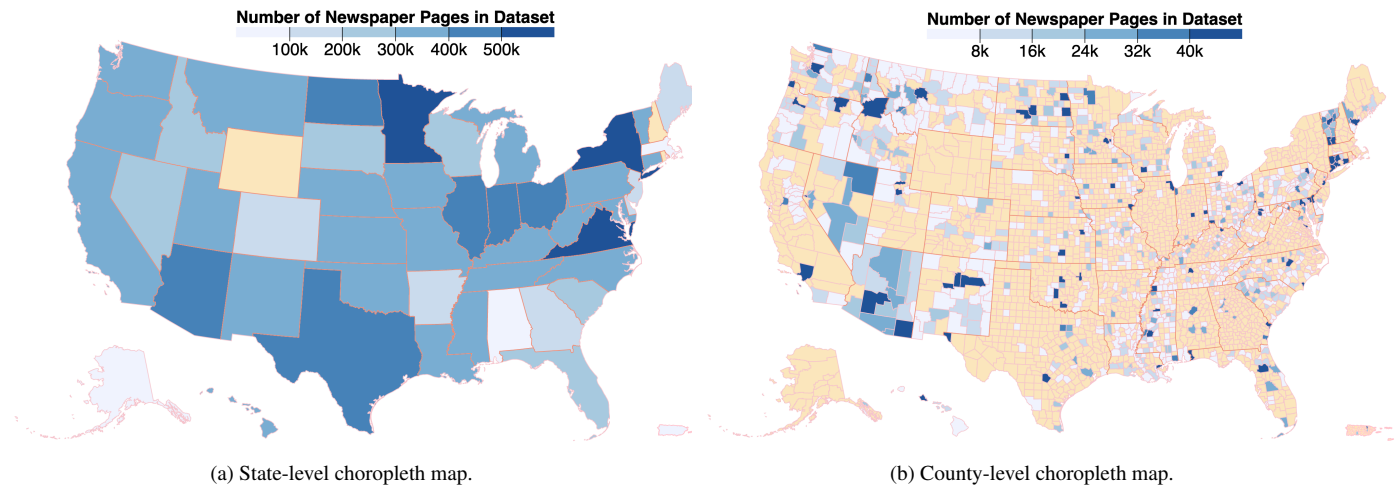(b) County-level choropleth map.

Figure 1: Choropleth maps at the state and county level showing the geographic coverage of the 16.3 million Chronicling America historic newspaper pages included in the Newspaper Navigator dataset. Yellow coloring indicates no pages cover the corresponding region. Puerto Rico is pictured in the bottom-right of each map.
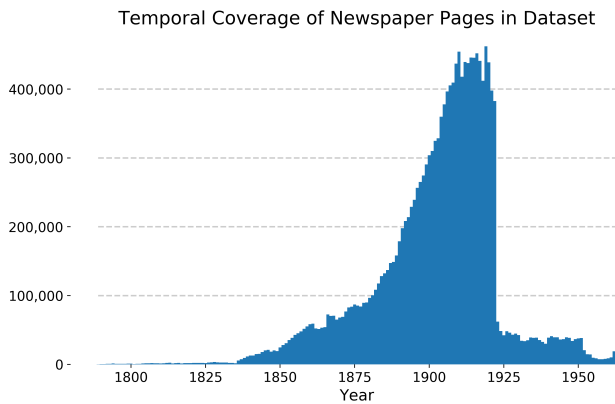

Figure 2: A histogram showing the temporal coverage of the 16.3 million Chronicling America historic newspaper pages included in the Newspaper Navigator dataset. A temporal cutoff of 1922 was used for Chronicling America newspaper digitization until 2016, explaining the corresponding dropoff in temporal coverage.

Visual Content Extraction
Other researchers have built tools and pipelines for extracting and analyzing visual content from historic documents, including newspapers, using deep learning.[3] PageNet utilizes a Fully Convolutional Network for pixel-wise page boundary extraction for historic documents [70]. dhSegment is a deep learning framework for historical document processing, including pixel-wise segmentation and extraction tasks [19]. Liebl and Burghardt benchmarked 11 different deep learning backbones for the pixel-wise segmentation of historic newspapers, including the separation of layout features such as text

and tables [41]. The AIDA collaboration at the University of Nebraska-Lincoln has applied deep learning techniques to newspaper corpora including Chronicling America and the Burney Collection of British Newspapers [44, 45, 46] for tasks such as poetic content recognition [47, 68], as well as visual content recognition using dhSegment [48]. Instead of a pixel-wise approach, we instead utilize bounding boxes, resulting in higher performance. In addition, our pipeline recognizes 7 different classes of visual content (headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements), extracts corresponding OCR, and generates image embeddings. Lastly, we deploy our visual content recognition pipeline at scale.

Article Disambiguation
Article disambiguation for historic newspaper pages has long been of interest to researchers. Groups that have studied this task include the IMPRESSO project [56], NewsEye project [57], and Google Newspaper Search [24].[4] Of particular note is the approach taken by Google Newspaper Search, which extracted headline blocks using OCR font size and area-perimeter ratio as features and utilized the extracted headlines to segment each page into individual articles [24].[5] We, too, focus on headline extraction because it serves as its own form of article disambiguation. However, unlike previous approaches, we treat headline extraction as a *visual* task at the image level, rather than a *textual* task at the OCR level. Our novel approach is to leverage the visual distinctiveness of headlines on the newspaper pages and train a classifier to predict bounding boxes around headlines on the page. The headline text within each bounding box is then extracted from the underlying METS/ALTO OCR.

---

[3]For approaches to historic document classification that do not utilize deep learning, see for example [40].

[4]Related work has focused on content segmentation for books [49].
[5]To our knowledge, the extraction and classification of visual content was outside of the scope of the project.

Lastly, it should be noted that proper article disambiguation requires the ability to filter out text from advertisements due to the ubiquity of advertisements. As with headlines, we treat advertisement identification as a visual task rather than a textual task because the advertisements are so naturally identified by their visual features. Because our visual content recognition model robustly identifies advertisements, we are able to disambiguate newspaper text from advertisement text.

Image Embeddings for Cultural Heritage Collections

In recent years, researchers have utilized image embeddings for visualizing and exploring visual content in cultural heritage collections. The Yale Digital Humanities Lab's PixPlot interface [31] and the National Neighbors project [43] utilize Inception v3 embeddings [69]. Google Arts & Culture's t-SNE Map utilizes embeddings produced by the Google search pipeline [29]. The Norwegian National Museum's Principal Components project [34] uses finetuned Caffe image embeddings [36]. Olivia Vane utilizes VGG-16 embeddings to visualizing the Royal Photographic Society Collection [72]. Likewise, Brian Foo has created a visualization of The American Museum of Natural History's image collection [32] using VGG-16 embeddings [62]. Refik Anadol uses embeddings to visualize the SALT Research collection [18]. Regarding visual content in historic newspapers in particular, Wevers and Smits have utilized Inception v3 embeddings to perform large-scale analysis of the CHRONIC and SIAMESET datasets derived from historic Dutch newspapers [74]. Their work includes the deployment of SIAMESE, a recommender system for advertisements in historic newspapers, as well as an analysis of training a new classification layer on top of the Inception embeddings to predict according to custom categories [74].

Indeed, in addition to supporting visualizations of latent spaces that capture semantic similarity, image embeddings are desirable for visual search and recommendation tasks due to the ability to perform fast similarity querying with them. Using ResNet-18 and ResNet-50 [35] models pre-trained on ImageNet, we generate image embeddings for the extracted visual content, which are included in the Newspaper Navigator dataset in order to support a range of visual search and recommendation tasks for the Chronicling America corpus.

CODE

All code discussed in this paper can be found in the public GitHub repository https://github.com/LibraryOfCongress/newspaper-navigator and is open source, placed in the public domain for unrestricted re-use. In addition, included in the repository are the finetuned visual content recognition model, the training set on which the model was finetuned, a Jupyter notebook for experimenting with the visual content recognition model, and a slideshow of predictions.

CONSTRUCTING THE TRAINING SET

Repurposing the Beyond Words Annotations

To create a training set for our visual content recognition model, we repurposed the publicly available annotations of photographs, illustrations, maps, comics, and editorial cartoons derived from Beyond Words, a crowdsourcing initiative

launched by the Library of Congress to engage the American public with the visual content in World War 1-era newspapers in Chronicling America. The Beyond Words platform itself was built using Scribe [14]. The crowdsourcing workflow consisted of three different tasks that volunteers could choose to perform:

1. *Mark*, in which users were asked to "draw a rectangle around each unmarked illustration or photograph excluding those in advertisements [and] enclose any caption or text describing the picture and the illustrator or photographer" [37].

2. *Transcribe*, in which users were asked to correct the OCR of the caption for each marked illustration or photograph, transcribe the author's name, and note the category ("Editorial Cartoon," "Comics/Cartoon," "Illustration," "Photograph," or "Map") [38].

3. *Verify*, in which users were asked to select the transcription of another volunteer that most closely matches the printed caption. Users were also able to filter out bad regions or provide their own transcriptions in the event that neither transcription was of good quality [39].

Up to 6 different individuals may have interacted with each annotation during this process. The annotation required achieving at least 51% agreement with volunteers at the *Transcribe* and *Verify* steps.

In order to finetune the visual content recognition model, it was first necessary to reformat the crowdsourced Beyond Words annotations into a proper data format for training a deep learning model. We chose the Common Objects in Context (COCO) dataset format [42], a standard data format for object detection, segmentation, and captioning tasks adopted by Facebook AI Research's Detectron2 deep learning platform for object detection [76].

The verified Beyond Words annotations used as training data were downloaded from the Beyond Words public website on December 1, 2019. To convert the JSON file available for download into a deep learning training set, we wrote a Python script to pull down the Chronicling America newspaper images utilized by Beyond Words and format the annotations according to the COCO standard. The script is available in the Newspaper Navigator GitHub repository.

We reiterate that the instructions for the "Mark" step asked users to "enclose any caption or text describing the picture and the illustrator or photographer" [37]; therefore, a model trained on these annotations learns to include relevant text within the bounding boxes for visual content, which can then be extracted from the corresponding METS/ALTO OCR in an automated fashion.

Adding Annotations

Because headlines and advertisements were not included in the Beyond Words workflow, we added annotations for headlines and advertisements for all images in the dataset. These annotations are not verified, as each page was annotated by only one person. In addition, due to the low number of annotated

Training/Validation Set Statistics	
Category	**Count**
Photograph	4,254
Illustration	1,048
Map	215
Comic/Cartoon	1,150
Editorial Cartoon	293
Headline	27,868
Advertisement	13,581
Total	48,409

Table 1: A table showing a breakdown of content for the 7 different classes in the training/validation dataset produced using the Beyond Words bounding box annotations, augmented with additional annotations.

Performance (Validation)		
Category	**AP**	**# in Val. Set**
Photograph	61.6%	879
Illustration	30.9%	206
Map	69.5%	34
Comic/Cartoon	65.6%	211
Editorial Cartoon	63.0%	54
Headline	74.3%	5,689
Advertisement	78.7%	2,858
Averaged (mAP)	63.4%	N/A
One Class	75.1%	9,931

Table 2: A table showing the average precision (AP) on validation data for the finetuned visual content recognition model on the different categories of content, as well as the number of instances of each category in the validation set. The *Averaged* row includes the mean average precision across the 7 classes. The *One Class* row is computed by combining all visual content into one class and computing average precision using the single class. This captures how much error is introduced by the detection of visual content versus the classification.

maps in the Beyond Words data (79 in total), we added annotations of 122 pages containing maps, which were retrieved by performing a keyword search of "map" on the Chronicling America search portal restricted to the years 1914-1918. We then downloaded the pages on which we identified maps, and we annotated all 7 categories of visual content on each page. Like the headline and advertisement annotations, these annotations are not verified.

Training Set Statistics

The augmented Beyond Words dataset in COCO format can be found in the Newspaper Navigator repository and is available for unrestricted re-use in the public domain. The dataset contains 3,559 World War 1-era Chronicling America pages with 48,409 annotations in total. The category breakdown of annotations appears in Table 1.

TRAINING THE VISUAL CONTENT RECOGNITION MODEL

To train a visual content recognition model for identifying the 7 classes of different newspaper content, we chose to finetune a pre-trained Faster-RCNN object detection model from Detectron2's Model Zoo using Detectron2 [76] and PyTorch [53]. Because model inference was the bottleneck on runtime in our pipeline, we chose the Faster-RCNN R50-FPN backbone, the fastest such backbone according to inference time. Though we could have utilized the highest performing Faster-RCNN backbone, which achieved approximately 5% higher mean average precision on the COCO [42] pre-training task at the expense of 2.5x the inference time, qualitative evaluation of predictions with the finetuned R50-FPN backbone indicated that the model was performing sufficiently for our purposes. Furthermore, we conjecture that the performance of our visual content recognition model is limited by noise in the training data, rather than model architecture and selection, for two reasons. First, the ground-truth Beyond Words labels were not complete because volunteers were only required to draw one bounding box per page (though more could be added). Second, there was non-trivial disagreement between Beyond Words annotators for the bounding box marking task due to

the heterogeneity of visual content layouts and the resulting ambiguities in the annotation task.[6]

All finetuning was performed using PyTorch [53] on a g4dn.2xlarge Amazon EC2 instance with a single NVIDIA T4 GPU. Finetuning the R50-FPN backbone was evaluated on a held-out validation set according to an 80%-20% split; the JSON files containing the training and validation splits are available for download with the GitHub repository. We used the following hyperparameters: a base learning rate of 0.00025, a batch size of 8, and 64 proposals per image. RESIZE_SHORTEST_EDGE and RANDOM_FLIP were utilized as data augmentation techniques.[7] Using early stopping, the model was finetuned for 77 epochs, which required 17 hours of runtime on the NVIDIA T4 GPU. The model weights file is publicly available and can be found in the GitHub repository for this project.

We report a mean average precision on the validation set of 63.4%; average precision for each category, as well as the number of validation instances in each category, can be found in Table 2. We chose average precision as our evaluation metric because it is the standard metric utilized in the computer vision community for benchmarking object detection tasks. Given a fixed intersection over union (IoU) threshold to evaluate whether a prediction is correct, average precision is computed by sorting all classifications according to prediction score, generating the corresponding precision-recall curve, and modifying it by drawing the smallest-area curve containing it that is also monotonically decreasing. According to the COCO

---

[6]In regard to the accuracy of the annotations, it is worth noting that Beyond Words was launched as an experiment; consequently, there were no interventions in workflow or community management after its launch, and the accuracy of the resulting annotations should be assessed accordingly.

[7]These are the only two data augmentation methods currently supported by Detectron2.
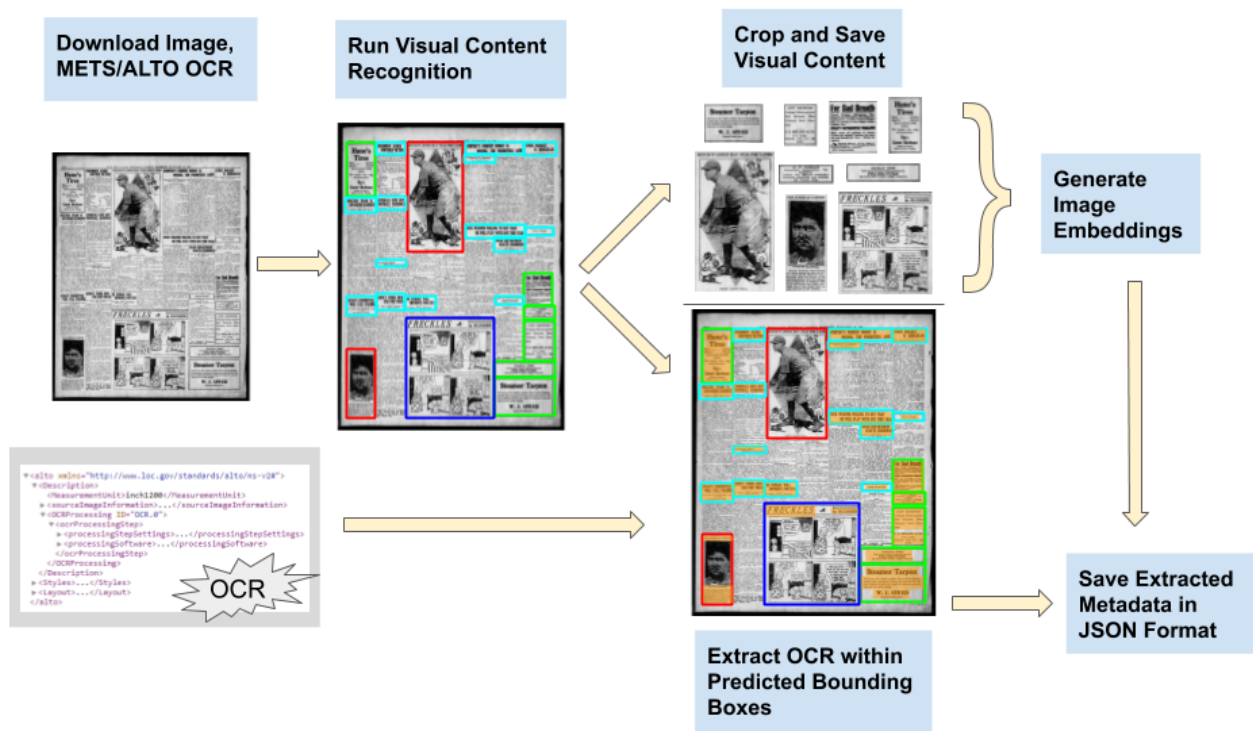
Figure 3: A diagram showing the steps of our pipeline.

standard, average precision is then computed by averaging the precision interpolated over 101 different recall values and 10 IoU thresholds from 50% to 95%. For our calculations, we utilized all predictions with confidence scores greater than 0.05 and discarded predictions with confidence scores below this threshold.[8]

THE PIPELINE

Building the Manifest

In order to create a full index of digitized pages for the pipeline to process, we used a forked version of the AIDA collaboration's `chronam-get-images` repository[9] to generate a manifest for each newspaper batch consisting of filepaths for each page in the batch.[10] Manifests consisting of 16,368,424 Chronicling America pages were compiled in total on March 17, 2020.

Steps of the Pipeline

In Figure 3, we present a diagram showing the pipeline workflow. Each manifest was processed in series by our pipeline. The pipeline code consists of six distinct steps:

---

[8] A confidence score of 0.05 is the default threshold cut for retaining predictions in Detectron2.

[9] https://github.com/bcglee/chronam-get-images

[10] More information on the batches can be found at https://chroniclingamerica.loc.gov/batches.

1. *Downloading the image and METS/ALTO XML for each page and downsampling the image by a factor of 6 to produce a lower resolution JPEG.* Downsampling was performed to reduce I/O and memory consumption, as well as to avoid the overhead introduced by the downsampling that Detectron2 would have to perform before each forward pass during model inference. This step was run in parallel across all 48 CPU cores on each EC2 instance. The files were pulled down directly from the Library of Congress's public AWS S3 buckets.

2. *Running the visual content recognition model inference on each image to produce bounding box predictions complete with coordinates, predicted classes, and confidence scores.* This step was run in parallel across all 4 GPUs on each EC2 instance. Predictions with confidence scores greater than 0.05 were saved. We chose to save predictions with low confidence scores in order to allow a user to select a threshold cut based on the user's ideal tradeoff between precision and recall.

3. *Extracting the OCR within each predicting bounding box.* This step required parsing the METS/ALTO XML and was run in parallel across all 48 CPU cores on each EC2 instance.

4. *Cropping and saving the extracted visual content as downsampled JPEGs (for all classes other than headlines).* This step was run in parallel across all 48 CPU cores on each EC2 instance.

5. *Generating ResNet-18 and ResNet-50 embeddings for the cropped and saved images with confidence scores of greater*

Newspaper Navigator Dataset Statistics

Category	Count ≥ Confidence Score		
	≥ 0.9	≥ 0.7	≥ 0.5
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 3: A table showing a breakdown of extracted content in the Newspaper Navigator dataset. Three different cuts on confidence score are presented to show the effect of the cut choice on the resulting dataset when favoring precision or recall.

than or equal to 0.5. This step was implemented using a forked version of img2vec[11] [59]. This step was run in parallel across all 4 GPUs on each EC2 instance. The ResNet-18 and ResNet-50 embeddings were extracted from the penultimate layer of each respective architecture after being trained on ImageNet (the models themselves were downloaded from `torchvision.models` in PyTorch [53]). The 2,048-dimensional ResNet-50 embeddings were selected due to ResNet-50's high performance and fast inference time relative to other image recognition models [21]. The 512-dimensional ResNet-18 embeddings were also generated due to their lower dimensionality, enabling faster computation for search and recommendation tasks.

6. *Saving the extracted metadata and cropped images.* The format of the saved metadata is described in the next section.

Running the Pipeline at Scale
All pipeline processing was performed on 2 g4dn.12xlarge Amazon AWS EC2 instances, each with 48 Intel Cascade Lake vCPUs and 4 NVIDIA T4 GPUs. All pipeline code was written in Python 3. In total, the pipeline successfully processed 16,368,041 pages (99.998%) in 19 days of wall-clock time. The manifests of pages that were successfully processed, as well as the 383 pages that failed, can be found in the Newspaper Navigator GitHub Repository.

THE NEWSPAPER NAVIGATOR DATASET

Statistics & Visualizations
A statistical breakdown of extracted content in the Newspaper Navigator dataset is presented in Table 3. Because the choice of threshold cut on confidence score affects the number of resulting visual content in the Newspaper Navigator dataset, we include statistics for three different threshold cuts of 0.5, 0.7, and 0.9.

---

[11]https://github.com/bcglee/img2vec

In Figure 4, we show visualizations of the number of photographs, illustrations, maps, comics, editorial cartoons, headlines, and advertisements in the Newspaper Navigator dataset according to year of publication. These visualizations show the average number of appearances per page of each of the seven classes over time, as well as the average fraction of the page covered by each of the seven classes over time. As in Table 3, we show three different cuts. With these visualizations, we can observe trends such as the rise of photographs at the turn of the 20th century and the gradual increase in the amount of page space covered by headlines from 1880 to 1920.

To demonstrate questions that we can begin to answer with the Newspaper Navigator dataset, we have included Figure 5, a visualization showing maps of the Civil War identified by searching the visual content for all 278,094 pages published between 1861 and 1865 in the dataset.[12] From this collection alone, researchers can study Civil War battles, the history of cartography, differences in print trends for northern and southern newspapers, and map reproduction patterns ("virality").
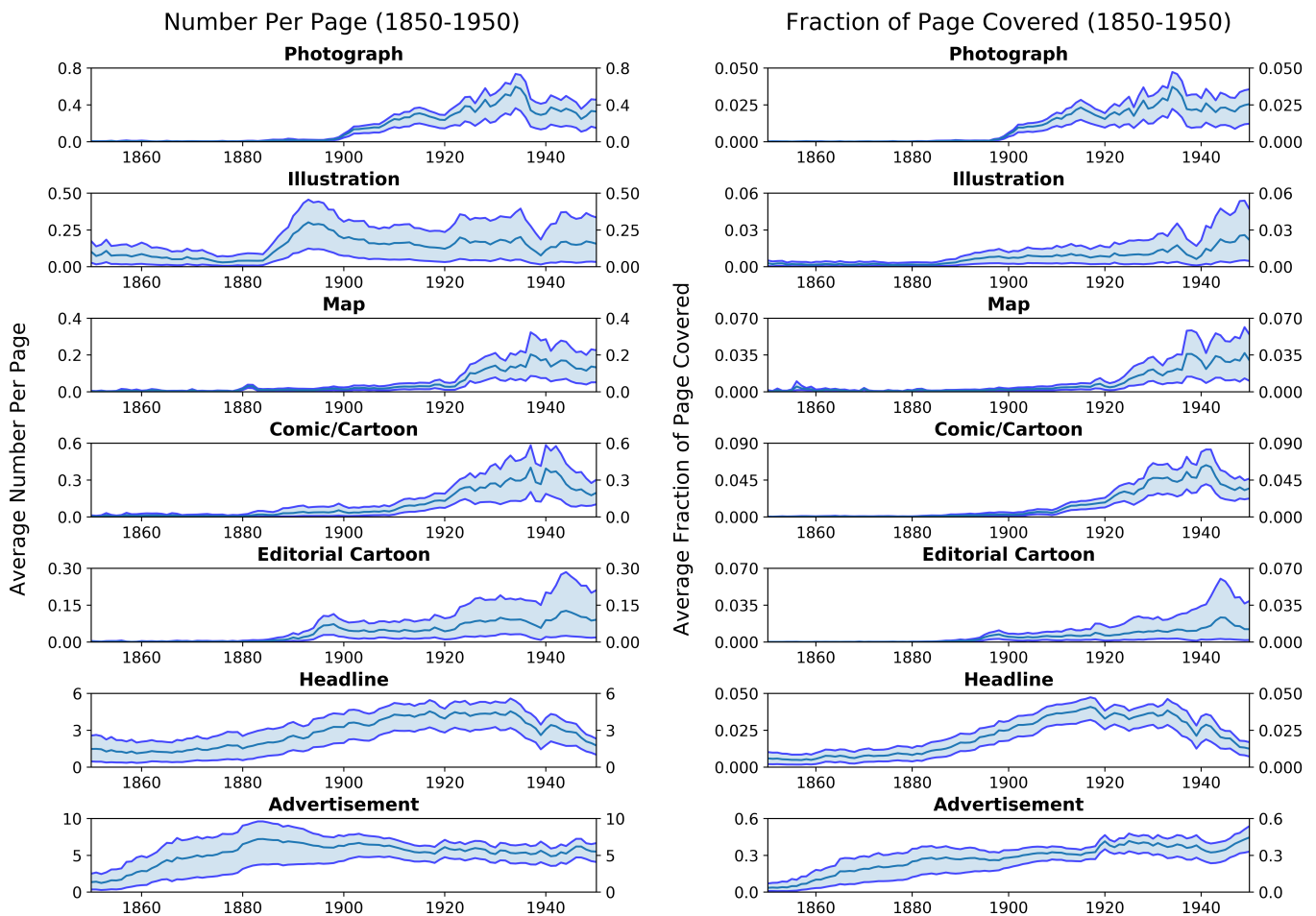
Dataset Access and Format
The Newspaper Navigator dataset can be accessed via the Newspaper Navigator GitHub repository. We introduce the data format below, but more detailed instructions for use can be found at the webpage. For each processed page, an associated JSON file contains the following metadata:

- `filepath` [`str`]: the path to the image, relative to the Chronicling America file structure.[13]

- `pub_date` [`str`]: the publication date of the page, in the format `YYYY−MM−DD`.

- `boxes` [`list:list`]: a list containing the normalized coordinates of predicted boxes, indexed according to $[x_1, y_1, x_2, y_2]$, where (x_1, y_1) is the top-left corner of the box relative to the standard image origin, and (x_2, y_2) is the bottom-right corner .

- `scores` [`list:float`]: a list containing the confidence score associated with each box (only predicted boxes with a confidence score ≥ 0.5 were retained).

- `pred_classes` [`list:int`]: a list containing the predicted class for each box using the following mapping of integers to classes:

    - 0 → Photograph
    - 1 → Illustration
    - 2 → Map
    - 3 → Comics/Cartoon
    - 4 → Editorial Cartoon
    - 5 → Headline
    - 6 → Advertisement

- `ocr` [`list:str`]: a list containing the OCR of white space-separated strings identified within each box.

---

[12]The visualization was created using [16].
[13]For example, see https://chroniclingamerica.loc.gov/data/batches/.

Number Per Page (1850-1950)   **Fraction of Page Covered (1850-1950)**

Average Number Per Page — panels: Photograph, Illustration, Map, Comic/Cartoon, Editorial Cartoon, Headline, Advertisement (x-axis: 1860, 1880, 1900, 1920, 1940)

Average Fraction of Page Covered — panels: Photograph, Illustration, Map, Comic/Cartoon, Editorial Cartoon, Headline, Advertisement (x-axis: 1860, 1880, 1900, 1920, 1940)

(a) A plot showing the average number of appearances per page of each of the seven classes of visual content, from 1850 to 1950.

(b) A plot showing the average fraction of each page covered by each of the seven classes of visual content, from 1850 to 1950.

Figure 4: Multipanel plots visualizing the number of photographs, illustrations, maps, comics, editorial cartoons, headlines, and advertisements in the Newspaper Navigator dataset, derived from 16.3 million historic newspaper pages in Chronicling America. **In each plot, the middle line corresponds to a cut of 0.7 on confidence score, and the upper and lower bounds of the confidence interval in light blue correspond to cuts of 0.5 and 0.9, respectively.** Note that the y-axis scales vary per category in both plots. The fraction of each page covered is included because it is a more consistent metric for complicated visual content layouts (such as photo montages and classified ads): predicted bounding boxes can vary greatly in number while still remaining correct and covering the same regions in aggregate.

- `visual_content_filepaths [list:str]`: a list containing the filepath for each cropped image (except headlines, which were not cropped and saved).

Another JSON file with the same file name with the suffix "`_embeddings`" includes the image embeddings in the following format; any prediction with a confidence score of less than 0.5 does not have a corresponding embedding:

- `filepath [str]`

- `resnet_50_embeddings [list:list]`: a list containing the 2,048-dimensional ResNet-50 embedding for each image (except headlines, for which embeddings were not generated).

- `resnet_18_embeddings [list:list]`: a list containing the 512-dimensional ResNet-18 embedding for each image (except headlines, for which embeddings were not generated).

- `visual_content_filepaths [list:list]`

Pre-packaged Datasets

In order to make the Newspaper Navigator dataset accessible to those without coding experience, we have also packaged smaller datasets derived from the Newspaper Navigator dataset that can be downloaded in bulk from our GitHub repository. These derivative datasets are grouped geographically and temporally and cover both visual content and textual content (machine-readable headlines, captions of visual content,

Performance for 19th Century Newspaper Pages

Category	AP (1850-1875)	AP (1875-1900)
Headline	21.2%	51.6%
Advertisement	7.3%	44.7%
Illustration	N/A	36.4%
One Class	12.1%	48.1%

Table 4: A table showing the average precision (AP) on test sets of 500 annotated pages from 1850 to 1875 and 500 annotated pages from 1875 to 1900. Due to the rarity of the other classes in the labeled data, only headlines, advertisements, and illustrations are included. As in Table 2, *One Class* refers to the average precision when combining all visual content into one class in order to capture how much error is introduced by the detection of visual content versus the classification.

etc.). One such example is the collection of Civil War maps shown in Figure 5. We will continue to add derivative datasets as Newspaper Navigator evolves.

DISCUSSION

Generalization to 19th Century Newspaper Pages

Given that the visual content recognition model has been trained on World War 1-era newspapers, it is natural to question the generalization ability of the model to 19th century newspapers. Though Figure 4 reveals trends consistent with intuition, such as the emergence of photographs in historic newspapers at the turn of the 20th century, it is still worthwhile to quantify generalization. To do so, we randomly selected 500 newspaper pages from 1850 to 1875 and 500 newspaper pages from 1875 to 1900 and annotated these pages. In Table 4, we present the average precision for headlines, advertisements, and illustrations in the test sets using our annotations as the ground truth (other classes were omitted due to their rarity in the annotated pages). Comparing the results in Table 4 to the results on the validation data in 2, we observe a moderate dropoff in performance for pages published between 1875 and 1900, as well as a more major dropoff in performance for pages published between 1850 and 1875. However, the extracted visual content from the pre-1875 pages in the Newspaper Navigator dataset is still of sufficient quality to enable novel analysis, as evidenced by the extracted collection of Civil War maps shown in Figure 5.

Partnering with Volunteer Crowdsourcing Initiatives

Our work is also a case study in partnering machine learning projects with volunteer crowdsourcing initiatives, a promising paradigm in which annotators are volunteers who learn about a new topic by participating. With the growing efforts of cultural heritage crowdsourcing initiatives such as the Library of Congress's By the People [1], Smithsonian's Digital Volunteers [15], the United States Holocaust Memorial Museum's History Unfolded [9], Zooniverse [63], the New York Public Library's Emigrant City [4], The British Library's LibCrowds [10], The Living with Machines project [11], and Trove's newspaper crowdsourcing initiative [20], there are more opportunities than ever to utilize crowdsourced data

for machine learning tasks relevant to cultural heritage, from handwriting recognition to botany taxonomic classification [60]. These partnerships also have the potential to provide insight into project design, decisions, workflows, and the context of the materials for which crowdsourcing contributions are sought. Along with Dielemann *et al*.'s work [30] training a neural network to classify galaxies using crowdsourced data from GalaxyZoo, we hope that our project encourages more machine learning researchers to partner with volunteer crowdsourcing projects, especially to study topics pertinent to cultural heritage.

CONCLUSION

In this paper, we have described our pipeline for extracting, categorizing, and captioning visual content, including headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements in historic newspapers. We present the Newspaper Navigator dataset, a dataset of these 7 types of extracted visual content from 16.3 million pages from Chronicling America. This is the largest dataset of its kind ever produced. In addition to releasing the Newspaper Navigator dataset, we have released our visual content recognition model for historic newspapers, as well as a new training dataset for this task based on annotations from Beyond Words, the Library of Congress Labs's crowdsourcing initiative for annotating and captioning visual content in World War 1-era newspapers in Chronicling America. All code has been placed in the public domain for unrestricted re-use.

FUTURE WORK

Future work on the pipeline itself includes improving the visual content recognition model's generalization ability for pre-20th century newspaper pages, especially for the 10.4% of the pages in the Newspaper Navigator dataset published before 1875. This could be accomplished by finetuning on a more diverse training set, which could be constructed by partnering with another volunteer crowdsourcing initiative such as the Living with Machines project [11]. One could also imagine training an ensemble of visual content recognition models on different date ranges. Given that only 10.4% of pages in the Newspaper Navigator dataset were published before 1875, it is straightforward to re-run the pipeline with an improved visual content recognition model on these pages.[14]

To improve the textual content extracted from the OCR, future work includes training an NLP pipeline to correct systematic OCR errors. In the second step of the Beyond Words pipeline, volunteers were asked to correct or enter the OCR that appears over each marked bounding box, resulting in approximately 10,000 corrected textual annotations to date. It is straightforward to construct training pairs of input and output in order to train a supervised model to correct OCR. Other approaches to OCR postprocessing include utilizing existing post-hoc OCR correction pipelines [17, 50, 51, 71], all of which could be benchmarked on the aforementioned Beyond Words training pairs.

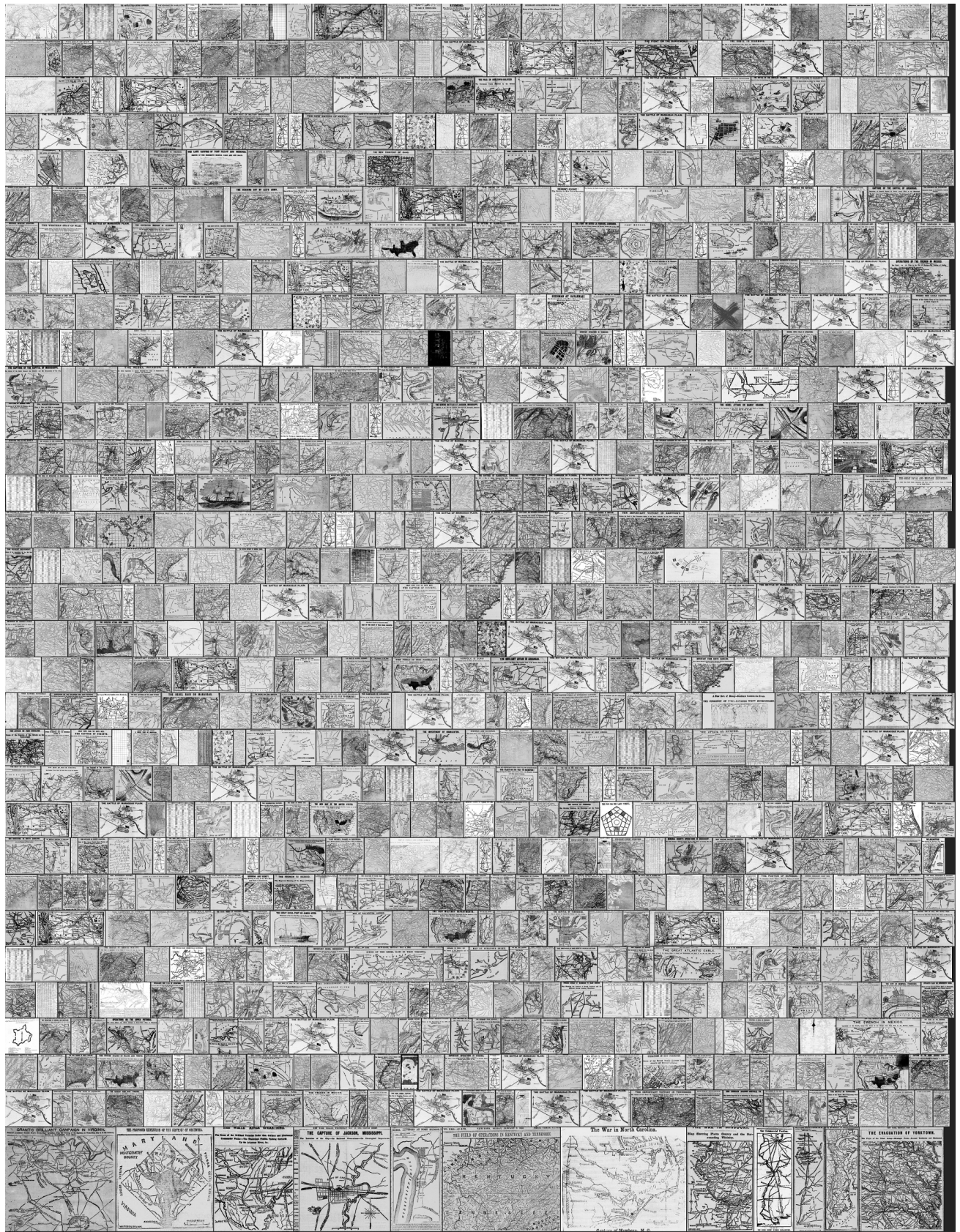[14]This simply requires replacing the model weights file and filtering the pages for processing by date range.

Figure 5: A visualization of the Civil War maps in the Newspaper Navigator dataset, filtered from 278,094 pages published from 1861 through 1865. Note that certain maps appear multiple times, indicating that they were reprinted. Only a small fraction of included images are false positives, suggesting the high performance of the visual content recognition model in the pipeline. This collection of Civil War maps is available as a pre-packaged dataset.

The future work that excites us most, however, consists of the many ways that the Newspaper Navigator dataset can be used. Our immediate future work consists of building a new search user interface called Newspaper Navigator that will be user tested in order to evaluate new methods of exploratory search. However, future work also includes investigating a range of digital humanities questions. For example, the Viral Texts [25, 28, 65, 64, 66, 77] and Oceanic Exchanges [27] projects have studied text reproduction patterns in 19th century newspapers, including newspapers in Chronicling America; the Newspaper Navigator dataset allows us to study photograph reproduction in 20th century newspapers. In addition, using the headlines in Newspaper Navigator, we can study which news cycles appeared in different regions of the United States at different times. These examples are just a few of many that we hope will be examined with the Newspaper Navigator dataset. We hope to inspire a wide range of digital humanities, public humanities, and creative computing projects.

ACKNOWLEDGMENTS

REFERENCES

[1] N/A. About By the People. `https://crowd.loc.gov/about/`. (N/A).

[2] N/A. About Chronicling America. (N/A). `https://chroniclingamerica.loc.gov/about/`

[3] N/A. About Delpher. (N/A). `https://www.delpher.nl/nl/platform/pages/helpitems?title=wat+is+delpher`

[4] N/A. About Emigrant City. `http://emigrantcity.nypl.org/#/about`. (N/A).

[5] N/A. About Newspapers.com. (N/A). `http://www.newspapers.com/about/` Library Catalog: www.newspapers.com.

[6] N/A. About The British Newspaper Archive. (N/A). `https://www.britishnewspaperarchive.co.uk/help/about`

[7] N/A. Civil War Maps - Newspaper and Current Periodical Reading Room (Serial and Government Publications Division, Library of Congress). (N/A). `https://www.loc.gov/rr/news/topics/civilwarmaps.html`

[8] N/A. Historical Newspapers âĂŞ BnL Open Data. (N/A). `https://data.bnl.lu/data/historical-newspapers/`

[9] N/A. History Unfolded: US Newspapers and the Holocaust. `https://newspapers.ushmm.org/about/project`. (N/A).

[10] N/A. LibCrowds Documentation. `https://docs.libcrowds.com/`. (N/A).

[11] N/A. Living with Machines: About Us. `https://livingwithmachines.ac.uk/about/`. (N/A).

[12] N/A. OurDigitalWorld: Digital Newspapers. (N/A). `https://ourdigitalworld.net/what-we-do/digital-newspapers/`

[13] N/A. Papers Past. `https://natlib.govt.nz/collections/a-z/papers-past`. (N/A).

[14] N/A. Scribe: Document Transcription, Crowdsourced. `https://scribeproject.github.io/`. (N/A).

[15] N/A. Smithsonian Digital Volunteers: Transcription Center: About. `https://transcription.si.edu/about`. (N/A).

[16] Dmitry Alimov. 2017. Collage Maker. (2017). `https://github.com/delimitry/collage_maker`

[17] Chantal Amrhein and Simon Clematide. 2018. Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods. *Journal for Language Technology and Computational Linguistics (JLCL)* 33, 1 (2018), 49–76. `https://doi.org/10.5167/uzh-162394`

[18] Refik Anadol. 2020. Archive Dreaming. `http://refikanadol.com/works/archive-dreaming/`. (2020).

[19] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhSegment: A generic deep-learning approach for document segmentation. In *Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on*. IEEE, 7–12.

[20] Marie-Louise Ayres. 2013. 'Singing for their supper': Trove, Australian newspapers, and the crowd. In *IFLA WLIC 2013*.

[21] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. 2018. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access* 6 (2018), 64270–64277. DOI: `http://dx.doi.org/10.1109/ACCESS.2018.2877890`

[22] Mazelan bin Anuar, Cally Law, and Soh Wai Yee. 2012. Challenges of Digitizing Vernacular Newspapers & Preliminary Study of User Behaviour on NewspaperSG's Multilingual UI. In *IFLA 2012 Pre-Conference: "The Electronic Re-evolution - News Media in the Digital Age"*. Mikkeli, Finland.

[23] Steve Cassidy. 2016. Publishing the Trove Newspaper Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), Portorož, Slovenia, 4520–4525. `https://www.aclweb.org/anthology/L16-1715`

[24] Krishnendu Chaudhury, Ankur Jain, Sriram Thirthala, Vivek Sahasranaman, Shobhit Saxena, and Selvam Mahalingam. 2009. Google Newspaper Search – Image Processing and Analysis Pipeline. In *2009 10th International Conference on Document Analysis and Recognition*. IEEE, Barcelona, Spain, 621–625. `DOI:` `http://dx.doi.org/10.1109/ICDAR.2009.272`

[25] Ryan Cordell. 2015. Reprinting, Circulation, and the Network Author in Antebellum Newspapers1. *American Literary History* 27, 3 (2015), 417–445. `DOI:` `http://dx.doi.org/10.1093/alh/ajv028`

[26] Ryan Cordell. 2017. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History* 20 (2017), 188 – 225.

[27] Ryan Cordell, M. H Beals, Isabel G Russell, Julianne Nyhan, Ernesto Priani, Marc Priewe, Hannu Salmi, Jaap Verheul, Raquel Alegre, Tessa Hauswedell, and et al. 2019. Oceanic Exchanges. (Oct 2019). `DOI:` `http://dx.doi.org/10.17605/OSF.IO/WA94S`

[28] Ryan Cordell and Abby Mullen. 2017. âĂIJFugitive VersesâĂİ: The Circulation of Poems in Nineteenth-Century American Newspapers. *American Periodicals: A Journal of History & Criticism* 27 (2017), 29 – 52.

[29] Cyril Diagne, Nicolas Barradeau, and Simon Doury. 2018. t-SNE Map. `https://experiments.withgoogle.com/t-sne-map`. (2018).

[30] Sander Dieleman, Kyle W. Willett, and Joni Dambre. 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society* 450, 2 (04 2015), 1441–1459. `DOI:` `http://dx.doi.org/10.1093/mnras/stv632`

[31] Douglas Duhaime. 2020. PixPlot. `https://github.com/YaleDHLab/pix-plot`. (2020).

[32] Brian Foo. 2020. Visualizing AMNH Image Collection with Machine Learning. `https://github.com/amnh-sciviz/image-collection`. (2020).

[33] T. GrÃijning, R. Labahn, M. Diem, F. Kleber, and S. Fiel. 2018. READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. 351–356. `DOI:` `http://dx.doi.org/10.1109/DAS.2018.38`

[34] Francoise Hanssen-Bauer, Magnus Bognerud, Dag Hensten, Gro Benedikte Pedersen, Even Westvang, and Audun Mathias ÃŸygard. 2018. t-SNE Map. `https://www.nasjonalmuseet.no/en/about-the-national-museum/collection-management---behind-the-scenes/digital-collection-management/project-principal-components/`. (2018).

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[36] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[37] LC Labs. 2017a. Beyond Words ("Mark"). (2017). `http://beyondwords.labs.loc.gov/#/mark`

[38] LC Labs. 2017b. Beyond Words ("Transcribe"). (2017). `http://beyondwords.labs.loc.gov/#/transcribe`

[39] LC Labs. 2017c. Beyond Words ("Verify"). (2017). `http://beyondwords.labs.loc.gov/#/verify`

[40] Benjamin Charles Germain Lee. 2018. 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 (11 2018), 513–535. `DOI:` `http://dx.doi.org/10.1093/llc/fqy063`

[41] Bernhard Liebl and Manuel Burghardt. 2020. An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers. (2020).

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[43] Matthew Lincoln, Golan Levin, Sarah Reiff Conell, and Lingdong Huang. 2019. National Neighbors: Distant Viewing the National Gallery of Art's Collection of Collections. `https://nga-neighbors.library.cmu.edu`. (2019).

[44] Elizabeth Lorang. 2018. Patterns, Collaboration, Practice: Algorithms as Editing for Historic Periodicals. (2018), 14.

[45] Elizabeth Lorang and Leen-Kiat Soh. 2019a. Application of the Image Analysis for Archival Discovery TeamâĂŹs First- Generation Methods and Software to the Burney Collection of British Newspapers. (2019), 21.

[46] Elizabeth Lorang and Leen-Kiat Soh. 2019b. Using Chronicling AmericaâĂŹs Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures. (2019), 19.

[47] Elizabeth Lorang, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. 2015. Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections. *D-Lib Mag.* 21 (2015).

[48] Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. 2020. Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project. (2020). `https://labs.loc.gov/static/labs/work/reports/UNL-final-report.pdf`

[49] Lara McConnaughey, Jennifer Dai, and David Bamman. 2017. The Labeled Segmentation of Printed Books. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 737–747. DOI:`http://dx.doi.org/10.18653/v1/D17-1077`

[50] Quoc-Dung Nguyen, Duc-Anh Le, and Ivan Zelinka. 2019. OCR Error Correction for Unconstrained Vietnamese Handwritten Text. In *Proceedings of the Tenth International Symposium on Information and Communication Technology (SoICT 2019)*. Association for Computing Machinery, New York, NY, USA, 132âĂŞ138. DOI: `http://dx.doi.org/10.1145/3368926.3369686`

[51] T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen, and A. Doucet. 2019. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 29–38. DOI:`http://dx.doi.org/10.1109/JCDL.2019.00015`

[52] Bob Nicholson. 2013. THE DIGITAL TURN: Exploring the methodological possibilities of digital newspaper archives. *Media History: Special Issue: Journalism and History: Dialogues* 19, 1 (2013), 59–73. `http://www.tandfonline.com/doi/abs/10.1080/13688804.2012.752963`

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`

[54] AleÅą PekÃąrek and Marieke Willems. 2012. The Europeana Newspapers âĂŞ A Gateway to European Newspapers Online. In *Progress in Cultural Heritage Preservation*, Marinos Ioannides, Dieter Fritsch, Johanna Leissner, Rob Davies, Fabio Remondino, and Rossella Caffo (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 654–659.

[55] National Digital Newspaper Program. 2020. Chronicling America Guidelines & Resources. (2020). `http://www.loc.gov/ndnp/guidelines/`

[56] Impresso Project. 2017. Impresso Project Overview. (2017). `https://impresso-project.ch/project/overview/`

[57] Juha Rautiainen. 2019. Opening Digitized Newspapers for Different User Groups - Successes and Challenges. In *IFLA WLIC 2019 - Athens, Greece - Libraries: Dialogue for Change*.

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99. `http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf`

[59] Christian Safka. 2019. img2vec. (Nov. 2019). `https://github.com/christiansafka/img2vec` original-date: 2017-09-21T08:19:17Z.

[60] Eric Schuettpelz, Paul B.ÂăFrandsen, Rebecca B.ÂăDikow, Abel Brown, Sylvia Orli, Melinda Peters, Adam Metallo, Vicki A.ÂăFunk, and Laurence J.ÂăDorr. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5 (2017). DOI: `http://dx.doi.org/10.3897/BDJ.5.e21139`

[61] Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. 2016. DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, Shenzhen, China, 471–476. DOI:`http://dx.doi.org/10.1109/ICFHR.2016.0093`

[62] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[63] Robert J. Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *WWW '14 Companion*.

[64] D. A. Smith, R. Cordell, and E. M. Dillon. 2013. Infectious texts: Modeling text reuse in nineteenth-century newspapers. In *2013 IEEE International Conference on Big Data*. 86–94. DOI: `http://dx.doi.org/10.1109/BigData.2013.6691675`

[65] D. A. Smith, R. Cordell, E. M. Dillon, N. Stramp, and J. Wilkerson. 2014. Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*. 183–192. DOI: `http://dx.doi.org/10.1109/JCDL.2014.6970166`

[66] David A. Smith, Ryan Cordell, and Abby Mullen. 2015. Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History* 27, 3 (06 2015), E1–E15. DOI: `http://dx.doi.org/10.1093/alh/ajv029`

[67] T. Smits and W.J. Faber. 2018. CHRONIC (Classified Historical Newspaper Images). `http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images`

[68] Leen-Kiat Soh, Elizabeth Lorang, and Yi Liu. 2018. Aida: Intelligent Image Analysis to Automatically Detect Poems in Digital Archives of Historic Newspapers. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 7837–7842. `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16880`

[69] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*. `http://arxiv.org/abs/1512.00567`

[70] Chris Tensmeyer, Brian Davis, Curtis Wigington, Iain Lee, and Bill Barrett. 2017. PageNet: Page Boundary Extraction in Historical Handwritten Documents. In *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (HIP2017)*. ACM, New York, NY, USA, 59–64. DOI: `http://dx.doi.org/10.1145/3151509.3151522`

[71] T. Underwood. 2017. Data Munging. `https://github.com/tedunderwood/DataMunging`. (2017).

[72] Olivia Vane. 2018. Visualising the Royal Photographic Society collection: Part 2. `https://www.vam.ac.uk/blog/digital/visualising-the-royal-photographic-society-collection-part-2`. (2018).

[73] M. Wevers and J. Lonij. 2017. SIAMESET. `http://lab.kb.nl/dataset/siameset`

[74] Melvin Wevers and Thomas Smits. 2019. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* 35, 1 (01 2019), 194–207. DOI: `http://dx.doi.org/10.1093/llc/fqy085`

[75] Marieke Willems and Rossitza Atanassova. 2015. Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights* 28, 1 (March 2015), 51–56. DOI: `http://dx.doi.org/10.1629/uksg.218` Number: 1 Publisher: UKSG in association with Ubiquity Press.

[76] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. `https://github.com/facebookresearch/detectron2`. (2019).

[77] Shaobin Xu, David Smith, Abigail Mullen, and Ryan Cordell. 2014. Detecting and Evaluating Local Text Reuse in Social Networks. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*. Association for Computational Linguistics, Baltimore, Maryland, 50–57. DOI: `http://dx.doi.org/10.3115/v1/W14-2707`