The Newspaper Navigator Dataset

Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America

Benjamin Charles Germain Lee*
bcgl@cs.washington.edu
University of Washington
Library of Congress

Jaime Mears
jame@loc.gov
LC Labs
Library of Congress

Eileen Jakeway
ejakeway@loc.gov
LC Labs
Library of Congress

Meghan Ferriter
mefe@loc.gov
LC Labs
Library of Congress

Chris Adams
cadams@loc.gov
IT Design & Development
Library of Congress

Nathan Yarasavage
nyarasavage@loc.gov
National Digital Newspaper Program
Library of Congress

Deborah Thomas
deth@loc.gov
National Digital Newspaper Program
Library of Congress

Kate Zwaard
kzwa@loc.gov
LC Labs & Digital Strategy
Library of Congress

Daniel S. Weld
weld@cs.washington.edu
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 American newspapers. Over 16 million pages have been digitized 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 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. 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.

*This work was completed while an Innovator-in-Residence at the Library of Congress, as well as a Ph.D. Student at the University of Washington.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10.
https://doi.org/10.1145/3340531.3412767

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Document structure**; • **Computing methodologies** → **Information extraction**; • **Applied computing** → **Digital libraries and archives**.

KEYWORDS

Information Retrieval; Document Analysis; Dataset; Historic Newspapers; Chronicling America; Newspaper Navigator; Digital Libraries and Archives; Digital Humanities; Public Domain

ACM Reference Format:
Benjamin Charles Germain Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. 2020. The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412767

1 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 in historic local newspapers. Over 16 million pages of newspapers published between 1789 to 1963 are publicly available online through a search portal and 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 [24, 51]. 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 [8]. 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 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 which 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 engage the American public and begin the construction of datasets of visual content within Chronicling America, the Library of Congress Labs launched the Beyond Words crowdsourcing initiative in 2017.[1] Volunteers were asked to draw bounding boxes around photographs, illustrations, comics, editorial cartoons, and maps in World War 1-era Chronicling America newspapers; they were also asked to transcribe captions by correcting the OCR within each bounding box, as well as record the content creator. Approximately 10,000 verified annotations have been collected to date.

Our research builds on Beyond Words 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 pretrained Faster-RCNN implementation from Detectron2's Model Zoo [54, 68]. 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. 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, to enable fast similarity querying for search and recommendation tasks, we generate image embeddings for the extracted visual content using ResNet 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 Newspaper 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. The dataset can be found at https://news-navigator.labs.loc.gov.

2 RELATED WORK

2.1 Corpora & Datasets

Over the past 15 years, efforts across the world to digitize historic newspapers have been remarkably successful [47]. In addition to Chronicling America, examples of large repositories of digitized newspapers include Trove [22], Europeana [50, 67], Delpher [3], The British Newspaper Archive [6], OurDigitalWorld [12], Papers Past [13], NewspaperSG [21], newspapers.com [5] and Google Newspaper Search [23]. These repositories have inspired the construction of datasets for related supervised learning tasks. In addition to Beyond Words, datasets for historic newspaper recognition include the National Library of Luxembourg's historic newspaper datasets [7] that include segmented articles and advertisements; KBK-1M, a dataset of 1,603,396 images with captions extracted from historic Dutch newspapers; CHRONIC, a dataset of 452,543 images in historic Dutch newspapers [60]; and SIAMESET, a dataset of 426,777 advertisements in historic Dutch newspapers [65]. Datasets for machine learning tasks with historical documents include READ-BAD [30] and DIVA-HisDB [57]. However, all of these datasets are subsets of visual content rather than comprehensive datasets of extracted content from full corpora. Our work uses 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 5.

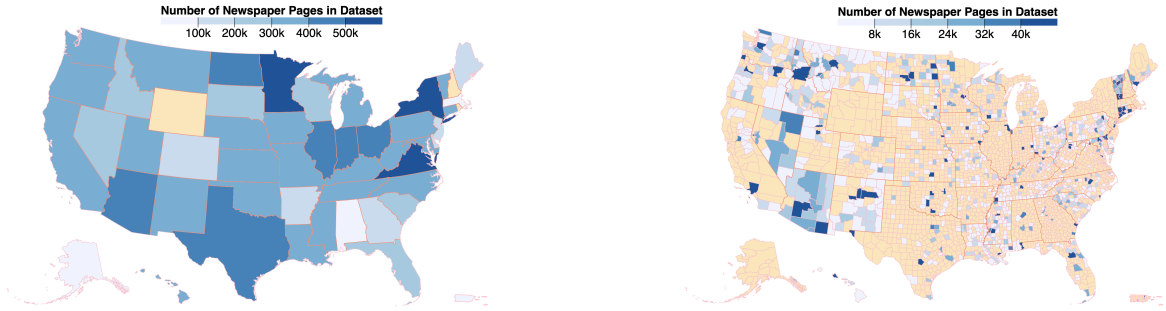

Figure 1: Choropleth maps at the state and county level showing the geographic coverage of the 16.3 million Chronicling America historic newspaper pages included in the Newspaper Navigator dataset. Yellow coloring indicates that no pages cover the corresponding region. Puerto Rico is pictured in the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

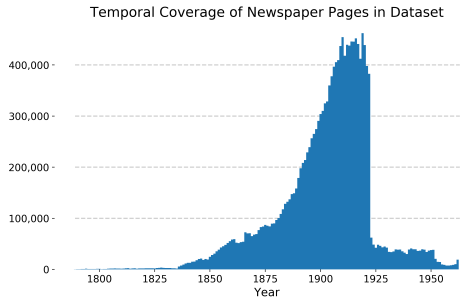

Figure 2: A histogram showing the temporal coverage of the 16.3 million Chronicling America historic newspaper pages included in the Newspaper Navigator dataset.

2.2 Visual Content Extraction

Other researchers have built tools and pipelines for extracting and analyzing visual content from historic documents using deep learning.[3] PageNet utilizes a Fully Convolutional Network for pixel-wise page boundary extraction for historic documents [63]. dhSegment is a deep learning framework for historical document processing, including pixel-wise segmentation and extraction tasks [18]. 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 [38]. The AIDA collaboration has applied deep learning techniques to newspaper corpora including Chronicling America and the Burney Collection of British Newspapers [41–43] for tasks such as poetic content recognition [44, 61] and visual content recognition using dhSegment [45]. Instead of a pixel-wise approach, we utilize bounding boxes, resulting in higher performance. In addition, our pipeline recognizes 7 different classes of visual content, extracts corresponding OCR, and generates image embeddings. Lastly, we deploy our visual content recognition pipeline at scale.

[3]For approaches to historic document classification that do not utilize deep learning, see for example [37].

2.3 Article Disambiguation

Article disambiguation for historic newspaper pages has long been of interest to researchers, including the IMPRESSO project [52], NewsEye project [53], and Google Newspaper Search [23]. 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 perform article segmentation [23].[4] 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 and train a classifier to predict bounding boxes around headlines on the page. The headline text within each bounding box is extracted from the METS/ALTO OCR.

Lastly, proper article disambiguation requires the ability to filter out text from advertisements due to their ubiquity. 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.

2.4 Image Embeddings and Cultural Heritage

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 [28] and the National Neighbors project [40] utilize Inception v3 embeddings [62]. Google Arts & Culture's t-SNE Map utilizes embeddings produced by the Google search pipeline [27]. The Norwegian National Museum's Principal Components project [31] uses finetuned Caffe image embeddings [33]. Olivia Vane utilizes VGG-16 embeddings to visualize the Royal Photographic Society Collection [64]. Likewise, Brian Foo has created a visualization of The American Museum of Natural History's image collection [29] using VGG-16 embeddings [58]. Refik Anadol uses embeddings to visualize the SALT Research collection [17]. Regarding visual content in historic newspapers in

[4]To our knowledge, the extraction and classification of visual content was outside of the scope of the project.

particular, Wevers and Smits utilize Inception v3 embeddings to analyze the CHRONIC and SIAMESET datasets described in Section 2.1. Their work includes deploying SIAMESE, a recommender system for historic newspaper advertisements, and analyzing the training of a new classification layer on top of the Inception embeddings to predict custom categories [66]. 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 [32] models pre-trained on ImageNet, we generate image embeddings for the extracted visual content, which are included in the Newspaper Navigator dataset.

3 CODE

All code can be found in the public GitHub repository: https://github.com/LibraryOfCongress/newspaper-navigator. All code 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.

4 CONSTRUCTING THE TRAINING SET

4.1 Repurposing 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. Built using Scribe [14], the crowdsourcing workflow consisted of three tasks:

(1) *Mark*: 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" [34].
(2) *Transcribe*: users were asked to correct the OCR of the caption for each marked box, transcribe the author's name, and note the category ("Editorial Cartoon," "Comics/Cartoon," "Illustration," "Photograph," "Map") [35].
(3) *Verify*: users were asked to select the transcription of another volunteer that most closely matched the printed caption. Users could also filter out bad regions or provide new transcriptions if none were of sufficient quality [36].

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, we first reformatted 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 [39], a standard data format for object detection, segmentation, and captioning tasks adopted by Facebook AI Research's Detectron2 deep learning platform for object detection [68]. The verified Beyond Words annotations used as training data were downloaded from the Beyond Words public website on December 1, 2019.

Table 1: A breakdown of content for the 7 different classes in the training/validation dataset composed of the Beyond Words annotations and additional annotations.

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

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" [34]; 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.

4.2 Adding Annotations

Because headlines and advertisements were not included in the Beyond Words workflow, we added annotations for these categories for all images in the dataset. These annotations are not verified, as each page was annotated by only one person. Due to the low number of annotated maps in the Beyond Words data (79 in total), we also annotated 122 pages containing maps, which we retrieved by performing a keyword search of "map" on the Chronicling America search portal restricted to 1914-1918. We downloaded the pages on which we identified maps and annotated all 7 categories of visual content on each page. Like the headline and advertisement annotations, these annotations are not verified.

4.3 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. It contains 3,559 World War 1-era Chronicling America pages with 48,409 annotations. The annotation category breakdown appears in Table 1.

5 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 [68] and PyTorch [48]. 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 ~5% higher mean average precision on the COCO [39] pre-training task at the expense of 2.5x the inference time, qualitative evaluation of

Table 2: 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. *Averaged* is the mean average precision (mAP) across the 7 classes. *One Class* refers to the average precision when combining all visual content into a single class, capturing how much error is introduced by the detection of visual content versus the classification.

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

predictions with the finetuned R50-FPN backbone indicated that it was performing sufficiently. Furthermore, we conjecture that the performance of our model is limited by noise in the training data, rather than model architecture and selection. 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 when marking bounding boxes due to the complexity of visual content layouts.[5]

All finetuning was performed using PyTorch 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 in 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.[6] Using early stopping, we finetuned the model for 77 epochs, requiring 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 (AP) for each category, as well as the number of validation instances in each category, are reported in Table 2. We chose AP because it is the standard metric in the computer vision community for benchmarking object detection tasks. Given a fixed intersection over union (IoU) threshold to evaluate if a prediction is correct, AP 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 monotonically decreasing. For the COCO standard,

[5]Beyond Words was launched as an experiment, without interventions in workflow or community management; the annotation accuracy should be assessed accordingly.
[6]These were the only supported data augmentation methods at the time of training.

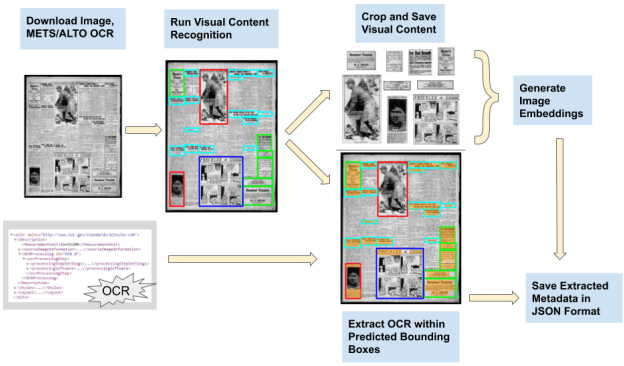

Figure 3: A diagram showing the steps of our pipeline.

AP 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, the default threshold in Detectron2.

6 THE PIPELINE

6.1 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[7] to generate a manifest of filepaths for each newspaper batch.[8] Manifests consisting of 16,368,424 Chronicling America pages were compiled in total on March 17, 2020.

6.2 Steps of the Pipeline

In Figure 3, we show the pipeline workflow. Each manifest was processed in series by our pipeline, which consists of 6 steps:

(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 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.

[7]https://github.com/bcglee/chronam-get-images
[8]For more information on the batches, see https://chroniclingamerica.loc.gov/batches.

Table 3: A breakdown of extracted content in the Newspaper Navigator dataset. Three cuts on confidence score are presented to show the effects when favoring precision or recall.

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

(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 than 0.05.* This step was implemented using a forked version of img2vec[9] [55]. 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.[10] 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 [20]. The 512-dimensional ResNet-18 embeddings were generated due to their lower dimensionality.
(6) *Saving the extracted metadata and cropped images.* The format of the saved metadata is described thoroughly on the dataset's landing page, https://news-navigator.labs.loc.gov.

6.3 Running the Pipeline at Scale

All pipeline processing was done on 2 g4dn.12xlarge Amazon AWS EC2 instances, each with 48 vCPUs (Intel Cascade Lake) and 4 NVIDIA T4 GPUs. All pipeline code was written in Python 3. The pipeline successfully processed 16,368,041 pages (99.998%) in 19 days of wall-clock time. The manifests of the processed pages, as well as the 383 pages that failed, are in our GitHub Repository.

7 THE NEWSPAPER NAVIGATOR DATASET

7.1 Statistics & Visualizations

A statistical breakdown of extracted content in the Newspaper Navigator dataset is presented in Table 3. Because the choice of cut on confidence score affects the cardinality of the resulting visual content, we include statistics for three different threshold cuts of 0.5, 0.7, and 0.9. 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, as well as the average fraction of the page covered, for each of the seven classes from 1850 to 1950. As in Table 3, we show three different cuts. In Figure 4, we 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.

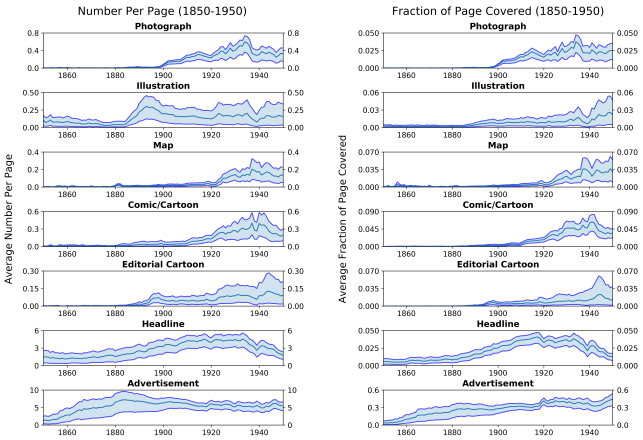

Figure 4: Multipanel plots visualizing the visual content in the Newspaper Navigator dataset over time (*left*: number per page; *right*: fraction of each page covered). In each plot, the middle line corresponds to a cut of 0.7 on confidence score, and the upper and lower bounds of the confidence interval in light blue correspond to cuts of 0.5 and 0.9, respectively.

7.2 Dataset Access

The Newspaper Navigator dataset can be accessed via the Newspaper Navigator GitHub repository, as well as the webpage https://news-navigator.labs.loc.gov/. This landing page contains a detailed descripton of the data format, as well as instructions for how to query the dataset. A search user interface is in development.

7.3 Pre-packaged Datasets

To make the Newspaper Navigator dataset accessible to those without coding experience, we have pre-packaged hundreds of smaller datasets as zip files, along with metadata in JSON and CSV formats. The pre-packaged datsets are grouped by year and visual content type, enabling users to download all of the 1921 headlines or 1864 maps, for example. Instructions for downloading the pre-packaged datasets can be found at the dataset landing page.

8 DISCUSSION

8.1 Generalization to 19th Century Newspapers

Given that the visual content recognition model has been trained on World War 1-era newspapers, it is natural to question how the model generalizes to 19th century newspapers. Though Figure 4 reveals trends consistent with intuition, such as the emergence of

[9] https://github.com/bcglee/img2vec
[10] We downloaded the pre-trained models from `torchvision.models` in PyTorch [48].

Table 4: Average precision (AP) on test sets of 500 annotated pages from 1850 to 1875 and 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 AP when combining all visual content into one class, capturing how much error is introduced by the detection of visual content versus the classification.

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%

photographs in historic newspapers around 1900, it is still worthwhile to quantify generalization. To do so, we randomly selected and annotated 500 newspaper pages from 1850 to 1875 and 500 pages from 1875 to 1900. In Table 4, we present the average precision for headlines, advertisements, and illustrations in the test sets using our annotations as the ground truth. Comparing the results to those in Table 2, we observe a moderate dropoff in performance for pages published between 1875 and 1900, as well as a more major dropoff for pages published between 1850 and 1875. However, the extracted visual content from these pages in the Newspaper Navigator dataset is still of sufficient quality for novel analysis.

8.2 Partnering with Volunteer Crowdsourcing

Our work is 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 [59], 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 [19], there are many opportunities to utilize crowdsourced data for machine learning tasks relevant to cultural heritage, from handwriting recognition to botany taxonomic classification [49, 56]. These partnerships have the potential to provide insight into project design, decisions, workflows, and the context of the materials for which crowdsourcing contributions are sought. We hope that our project encourages more machine learning researchers to partner with volunteer crowdsourcing projects, especially on topics pertinent to cultural heritage.

9 CONCLUSION & FUTURE WORK

We have described our pipeline for extracting, categorizing, and captioning visual content in historic newspapers, including headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements. We present the Newspaper Navigator dataset containing 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 this 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 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. 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 this subset.

To improve the extracted OCR, future work includes training a pipeline to correct systematic errors. In the second step of the Beyond Words pipeline, volunteers corrected the OCR appearing in 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 [16, 46], which could be benchmarked on the Beyond Words training pairs.

The future work that excites us most consists of the many ways that the Newspaper Navigator dataset can be used. We are currently building a new search user interface that will be user tested to evaluate new methods of exploratory search. Future work also includes investigating a range of digital humanities questions. For example, the Viral Texts [26] and Oceanic Exchanges [25] projects have studied text reproduction patterns in 19th century newspapers; 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 and when. These examples are just a few of many to be explored with the Newspaper Navigator dataset. We hope to inspire a wide range of digital humanities, public humanities, and creative computing projects.

ACKNOWLEDGMENTS

We thank Laurie Allen, Leah Weinryb Grohsgal, Abbey Potter, Robin Butterhof, Tong Wang, Mark Sweeney, and the entire National Digital Newspaper Program staff at the Library of Congress; Molly Hardy at the National Endowment for the Humanities; Stephen Portillo, Daniel Gordon, and Tim Dettmers at the University of Washington; Michael Haley Goldman, Eric Schmalz, and Elliott Wrenn at the United States Holocaust Memorial Museum; and Gabriel Pizzorno at Harvard University for their invaluable advice with this project. Lastly, we thank everyone who has contributed to Chronicling America and Beyond Words, without whom none of this work would be possible. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1762114, the Library of Congress Innovator-in-Residence Position, and the WRF/Cable Professorship.

REFERENCES

[1] N/A. About By the People. https://crowd.loc.gov/about/.
[2] N/A. About Chronicling America. https://chroniclingamerica.loc.gov/about/
[3] N/A. About Delpher. https://www.delpher.nl/nl/platform/pages/helpitems?title=wat+is+delpher
[4] N/A. About Emigrant City. http://emigrantcity.nypl.org/#/about.
[5] N/A. About Newspapers.com. http://www.newspapers.com/about/
[6] N/A. About The British Newspaper Archive. http://www.britishnewspaperarchive.co.uk/help/about
[7] N/A. BnL Historical Newspapers. https://data.bnl.lu/data/historical-newspapers/
[8] N/A. Civil War Maps - Newspaper and Current Periodical Reading Room (Library of Congress). https://www.loc.gov/rr/news/topics/civilwarmaps.html
[9] N/A. History Unfolded: US Newspapers and the Holocaust. https://newspapers.ushmm.org/about/project.
[10] N/A. LibCrowds Documentation. https://docs.libcrowds.com/.
[11] N/A. Living with Machines: About Us. https://livingwithmachines.ac.uk/about/.
[12] N/A. OurDigitalWorld: Digital Newspapers. https://ourdigitalworld.net/what-we-do/digital-newspapers/
[13] N/A. Papers Past. https://natlib.govt.nz/collections/a-z/papers-past.
[14] N/A. Scribe. https://scribeproject.github.io/.
[15] N/A. Smithsonian Digital Volunteers: About. https://transcription.si.edu/about.
[16] Chantal Amrhein and Simon Clematide. 2018. Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods. *JLCL* 33, 1 (2018), 49–76. https://doi.org/10.5167/uzh-162394
[17] Refik Anadol. 2020. Archive Dreaming. http://refikanadol.com/works/archive-dreaming/.
[18] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. 2018. dhSegment: A generic deep-learning approach for document segmentation. In *ICFHR '18*. IEEE.
[19] Marie-Louise Ayres. 2013. 'Singing for their supper': Trove, Australian newspapers, and the crowd. In *IFLA WLIC 2013*.
[20] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. 2018. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access* 6 (2018), 64270–64277. https://doi.org/10.1109/ACCESS.2018.2877890
[21] 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*. Mikkeli, Finland.
[22] Steve Cassidy. 2016. Publishing the Trove Newspaper Corpus. In *LREC'16*. Portorož, Slovenia, 4520–4525. https://www.aclweb.org/anthology/L16-1715
[23] Krishnendu Chaudhury, Ankur Jain, Sriram Thirthala, Vivek Sahasranaman, Shobhit Saxena, and Selvam Mahalingam. 2009. Google Newspaper Search: Image Processing and Analysis Pipeline. In *ICDAR '09*. IEEE, Barcelona, Spain, 621–625. https://doi.org/10.1109/ICDAR.2009.272
[24] Ryan Cordell. 2017. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History* 20 (2017), 188 – 225.
[25] 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. https://doi.org/10.17605/OSF.IO/WA94S
[26] Ryan Cordell and David Smith. 2017. Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines. (2017). http://viraltexts.org
[27] Cyril Diagne, Nicolas Barradeau, and Simon Doury. 2018. t-SNE Map. https://experiments.withgoogle.com/t-sne-map.
[28] Douglas Duhaime. 2020. PixPlot. https://github.com/YaleDHLab/pix-plot.
[29] Brian Foo. 2020. Visualizing AMNH Image Collection with Machine Learning. https://github.com/amnh-sciviz/image-collection.
[30] T. Grüning, 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 *13th IAPR Intl. Workshop on Document Analysis Systems*. 351–356. https://doi.org/10.1109/DAS.2018.38
[31] Francoise Hanssen-Bauer, Magnus Bognerud, Dag Hensten, Gro Benedikte Pedersen, Even Westvang, and Audun Mathias Øygard. 2018. t-SNE Map. https://github.com/nasjonalmuseet/propinquity.
[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).
[33] 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).
[34] LC Labs. 2017. Beyond Words, Mark. beyondwords.labs.loc.gov/#/mark
[35] LC Labs. 2017. Beyond Words, Transcribe. beyondwords.labs.loc.gov/#/transcribe
[36] LC Labs. 2017. Beyond Words, Verify. beyondwords.labs.loc.gov/#/verify
[37] Benjamin C.G. 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 (2018), 513–535. https://doi.org/10.1093/llc/fqy063
[38] Bernhard Liebl and Manuel Burghardt. 2020. An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers. arXiv:2004.07317 [cs.CV]
[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, and et al. 2014. Microsoft COCO: Common Objects in Context. In *ECCV '14*. Springer Intl. Publishing, 740–755.
[40] 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.
[41] Elizabeth Lorang. 2018. Patterns, Collaboration, Practice: Algorithms as Editing for Historic Periodicals. (2018).
[42] Elizabeth Lorang and Leen-Kiat Soh. 2019. Application of the Image Analysis for Archival Discovery Team's First- Generation Methods and Software to the Burney Collection of British Newspapers. (2019).
[43] Elizabeth Lorang and Leen-Kiat Soh. 2019. Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures. (2019).
[44] 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).
[45] Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. 2020. Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project. https://labs.loc.gov/static/labs/work/reports/UNL-final-report.pdf
[46] T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen, and A. Doucet. 2019. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *JCDL '19*. 29–38. https://doi.org/10.1109/JCDL.2019.00015
[47] 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
[48] Adam Paszke, Sam Gross, Francisco Massa, and et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NIPS '19*. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[49] Katelin D Pearson, Gil Nelson, Myla F J Aronson, and et al. 2020. Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience* (2020). https://doi.org/10.1093/biosci/biaa044
[50] Aleš Pekárek and Marieke Willems. 2012. The Europeana Newspapers: A Gateway to European Newspapers Online. In *Progress in Cultural Heritage Preservation*. Springer, Berlin, 654–659.
[51] National Digital Newspaper Program. 2020. Chronicling America Guidelines & Resources. http://www.loc.gov/ndnp/guidelines/
[52] Impresso Project. 2017. Impresso Project. impresso-project.ch/project/overview/
[53] Juha Rautiainen. 2019. Opening Digitized Newspapers for Different User Groups - Successes and Challenges. In *IFLA WLIC 2019*. Athens, Greece.
[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS '15*. 91–99. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf
[55] Christian Safka. 2019. img2vec. https://github.com/christiansafka/img2vec
[56] Eric Schuettpelz, Paul B. Frandsen, Rebecca B. Dikow, Abel Brown, Sylvia Orli, Melinda Peters, Adam Metallo, Vicki A. Funk, and Laurence J. Dorr. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5 (2017). https://doi.org/10.3897/BDJ.5.e21139
[57] 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 *ICFHR '16*. IEEE, Shenzhen, China, 471–476. https://doi.org/10.1109/ICFHR.2016.0093
[58] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR '15*.
[59] Robert J. Simpson, Kevin R. Page, and David De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. In *WWW '14 Companion*.
[60] T. Smits and W.J. Faber. 2018. CHRONIC (Classified Historical Newspaper Images). http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images
[61] Leen-Kiat Soh, Elizabeth Lorang, and Yi Liu. 2018. Aida: Intelligent Image Analysis to Automatically Detect Poems in Digital Archives of Historic Newspapers. In *IAAI '18*. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16880
[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR '16*. 2818–2826.
[63] 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 Intl. Workshop on Historical Document Imaging & Processing* (Kyoto, Japan). ACM, 59–64. https://doi.org/10.1145/3151509.3151522
[64] 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.
[65] M. Wevers and J. Lonij. 2017. SIAMESET. http://lab.kb.nl/dataset/siameset
[66] Melvin Wevers and Thomas Smits. 2019. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* 35, 1 (2019), 194–207. https://doi.org/10.1093/llc/fqy085
[67] Marieke Willems and Rossitza Atanassova. 2015. Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights* 28 (2015), 51–56. https://doi.org/10.1629/uksg.218
[68] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.