Digital Libraries, Intelligent Data Analytics, and Augmented Description:

A Demonstration Project

Final Report

Prepared by Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack

University Libraries & Department of Computer Science and Engineering

University of Nebraska–Lincoln

Submitted to the Library of Congress

January 10, 2020
Rev. March 26, 2020[1]
Rev. June 15, 2020[2]

[1] This revision included updated information about the code repository in section 4 (Code & Data), a new footnote about Beyond Words on page 10, and two additional tables (Table 9 and Table 10), and it corrected several typos.

[2] This revision removes two appendices, which featured work-in-progress reports completed throughout the project and slides from presentations delivered to the Library of Congress over the course of the project, in order to manage the scope and size of this final report for distribution and ease of use. The Table of Contents has been updated accordingly. Work-in-progress reports and slides from presentations are now available as stand-alone documents and are available from the Library of Congress and via projectaida.org. This revision also adds a final sentence to the Introduction, directing readers to the Discussion and Recommendations sections, based on their interests. It also corrects section numbers in the body of the document for subsections in 7: Discussion.

SUMMARY

From July 16-to November 8, 2019, the Aida digital libraries research team at the University of Nebraska-Lincoln collaborated with the Library of Congress on "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project." This demonstration project sought to (1) develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery; (2) understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections; and (3) enable the Library of Congress to plan for future possibilities. In pursuit of these goals, we focused our work around two areas: extracting and foregrounding visual content from Chronicling America (chroniclingamerica.loc.gov) and applying a series of image processing and machine learning methods to minimally processed manuscript collections featured in By the People (crowd.loc.gov). We undertook a series of explorations and investigated a range of issues and challenges related to machine learning and the Library's collections.

This final report details the explorations, addresses social and technical challenges with regard to the explorations and that are critical context for the development of machine learning in the cultural heritage sector, and makes several recommendations to the Library of Congress as it plans for future possibilities. We propose two top-level recommendations. First, the Library should focus the weight of its machine learning efforts and energies on social and technical infrastructures for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries. Second, we recommend that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections. Both of these top-level recommendations map to the three goals of the Library's 2019 digital strategy.

Within each top-level recommendation, we offer three more concrete, short- and medium-term recommendations. They include, under social and technical infrastructures: (1) Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage. (2) Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to the larger cultural heritage and other research communities. (3) Focus efforts on developing ground truth sets and benchmarking data and making these easily available. Nested under the recommendation to support ongoing explorations and investigations, we recommend that the Library: (4) Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate. (5) Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas. (6) Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations. Each of these recommendations speak to the investigation and challenge areas identified by Thomas Padilla in *Responsible Operations: Data Science, Machine Learning, and AI in Libraries.*

This demonstration project—via its explorations, discussion, and recommendations—shows the potential of machine learning toward a variety of goals and use cases, and it argues that the technology itself will not be the hardest part of this work. The hardest part will be the myriad challenges to undertaking this work in ways that are socially and culturally responsible, while also upholding responsibility to make the Library of Congress's materials available in timely and accessible ways. Fortunately, the Library of Congress is in a remarkable position to advance machine learning for cultural heritage organizations, through its size, the diversity of its collections, and its commitment to digital strategy.

TABLE OF CONTENTS

TABLES, FIGURES & CHARTS

TABLES

FIGURES

CHARTS

1 Introduction

In response to notice ID 030ADV19Q0274, "The Library of Congress – Pre-processing Pilot," the Aida digital libraries research team at the University of Nebraska-Lincoln (UNL) proposed "Digital Libraries, Intelligent Data Analytics, and Augmented Description: A Demonstration Project." The proposal was awarded a research services contract from the Library of Congress. From July 16-to November 8, 2019, members of the Aida research team conducted a series of explorations and analyses to assist the Library of Congress in assessing possible applications of machine learning within the Library. Three broad goals framed this work: (1) develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery; (2) understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections; and (3) enable the Library of Congress to plan for future possibilities.

This report summarizes the design of the demonstration project and the activities of the Aida research team; presents key findings resulting from these activities; makes recommendations for possible paths forward; and provides documentation for the major activities, including code, data, and reports-in-progress completed during the project. This report serves two main purposes: to document the work completed and to extrapolate from that work to broader implications for machine learning endeavors at the Library of Congress. Readers most interested in the larger social and technical implications of this work may wish to skip to Section 7: Discussion and Section 8: Recommendations.

2 Participants & Roles

University of Nebraska-Lincoln

- Elizabeth Lorang, senior adviser
- Leen-Kiat Soh, senior adviser
- Yi Liu, research associate and developer
- Chulwoo (Mike) Pack, research associate and developer
- Ashlyn Stewart, research assistant

Library of Congress[‡]

- Meghan Ferriter, Chief (Acting) LC Labs/Senior Innovation Specialist
- Abbey Potter, Senior Innovation Specialist
- Jaime Mears, Senior Innovation Specialist
- Eileen Jakeway, Innovation Specialist
- Tong Wang, Senior IT Specialist, OCIO
- Lauren Algee, Senior Innovation Specialist
- Victoria Van Hyning, Senior Innovation Specialist

3 Timeline

July 16, 2019

Project kick-off meeting held at the Library of Congress

[‡] In addition to these key contributors, many others at the Library of Congress supported this demonstration project in a variety of ways, including through their hospitality, encouragement, brainstorming, and interest in this project. We are grateful for and indebted to their efforts.

July 19-August 23, 2019

First-round of iterative development, onsite at the Library of Congress

August 26-November 8, 2019

Second round of iterative development, offsite at the University of Nebraska-Lincoln

November 6, 2019

Delivery of preliminary results via virtual meeting

November 7 – January 9

Development of open repository of code, data, and documentation; development of final report

January 10, 2020

Delivery of final results via in-person meeting at Library of Congress

4 Code & Data

Code and descriptions of data are available via the Library of Congress's GitHub organization page at the "Exploring ML with Project Aida" repository, https://github.com/LibraryOfCongress/Exploring-ML-with-Project-Aida. Following submission to the Library of Congress, code, data, and this report will also be available via projectaida.org.

5 Demonstration Project Design & Approach

With the size of the Library of Congress's digital collections and the many potential areas of impact, we might have pursued any number of questions in this demonstration project. Scoping our work, both with regard to the questions we pursued and the number and type of explorations, was critical. We anchored our work around two areas: (1) extracting and foregrounding visual content from Chronicling America (chroniclingamerica.loc.gov) through a variety of techniques and approaches and (2) applying a series of image processing and machine learning methods and techniques to minimally processed manuscript collections featured in By the People (crowd.loc.gov). We identified these areas of focus because they drew on collections already deemed significant by the Library of Congress and because they had a degree of ground-truthing work already completed. In addition, they offered the opportunity to explore the advantages and disadvantages and the strengths and weaknesses of computational/machine learning approaches as compared to data and information generated by experts, casual users, and researchers. Working with these collections had the further benefit of significant opportunity to create new, rich, and varied metadata about them, so that the Library might explore the ways in which more robust metadata might allow for alternative points of entry into the materials and the opportunity for Library staff and researchers to pursue questions of varying nature.

Ultimately, we designed a series of explorations that allowed us to investigate a range of issues and challenges related to machine learning and the Library's collections. The explorations were developed through an iterative process and in regular consultation with members of the Library of Congress staff, both to learn from their expertise and to make sure the questions we were pursuing were of value and interest to the Library. Through that process, some explorations merged, others concluded more quickly than others, and areas of inquiry seeded in one exploration began to sprout in others as well. Individually, the explorations pursued particular technical and collections-oriented questions. We also used the explorations as points of entry into—and paths to reflection about—larger issues, questions, and challenges for machine learning and cultural heritage.

6 THE EXPLORATIONS

This section presents an overview and details of six explorations: Document Segmentation; Graphic Element Classification and Text Extraction; Document Type Classification; Digitization Type Differentiation; Document Image Quality Assessment and Advanced Document Image Quality Assessment; and Document Clustering. Figure 1 and Table 1 identify and show relationships among the explorations and summarize them. In our look at each exploration, we identify guiding questions; outline and describe our approaches, techniques, and methods; present high-level results and analysis; and offer ideas toward future development and/or potential applications.

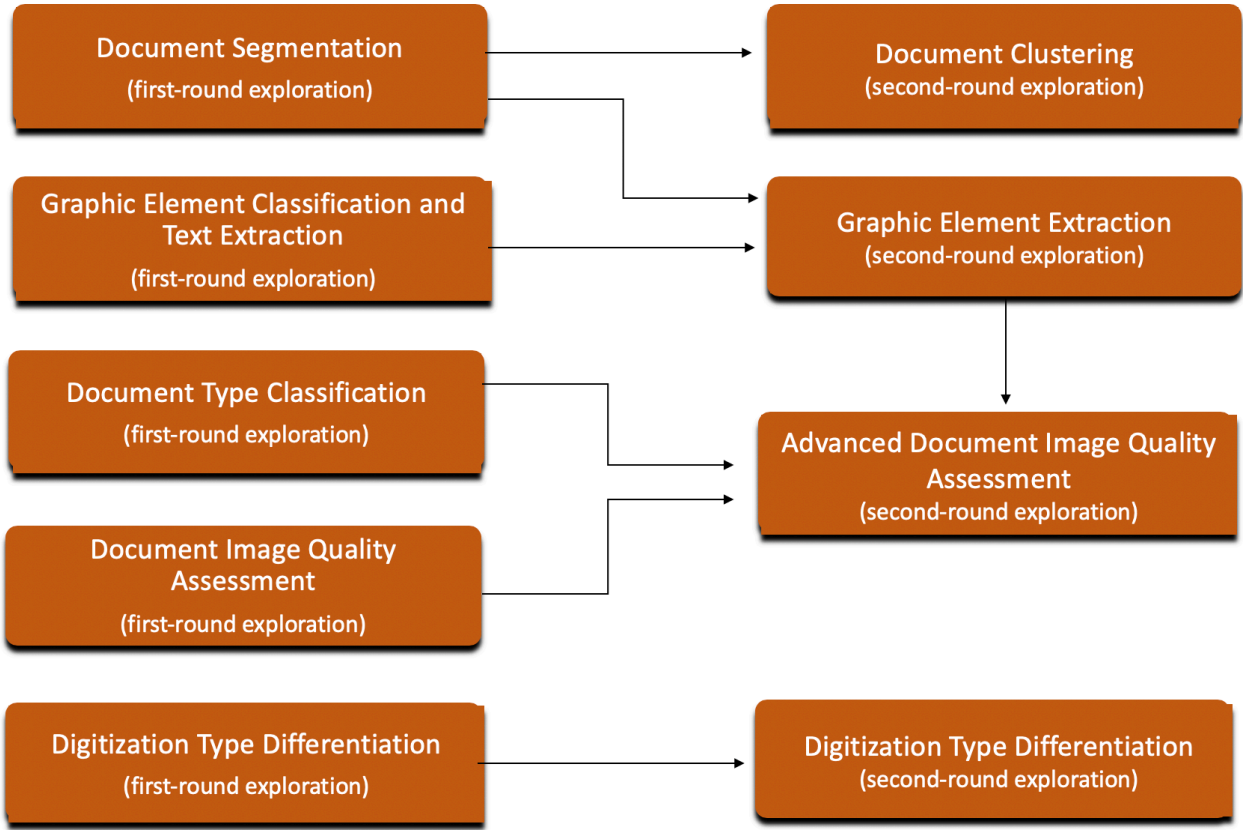


FIGURE 1. VISUAL REPRESENTATION OF THE EXPLORATIONS AND THEIR RELATIONSHIPS TO ONE ANOTHER.

TABLE 1. THE EXPLORATIONS PURSUED AS PART OF THE DEMONSTRATION PROJECT AND THEIR SELECTED POTENTIAL APPLICATIONS.

		Selected Potential Applications					
	Technical Objective	**Metadata generation (structural, descriptive, etc.)**	**Graphical content extraction**	**Influence decision-making for human and/or machine processing**	**Faceted data for end-users/researchers in search/discovery interface**	**Ground truth and benchmark sets for machine learning and image analysis projects competitions**	**Understanding collections**
First-Round Explorations							
Document Segmentation	Find and localize image-like components in newspaper pages	✓	✓		✓	✓	
Graphic Element Classification and Text Extraction	Find and localize graphical content, extract text from this content in newspapers	✓	✓		✓	✓	
Document Type Classification	Classify manuscript collection pages as handwritten, printed, mixed	✓		✓	✓	✓	✓
Document Image Quality Assessment	Analyze quality of manuscript collection page images	✓		✓	✓	✓	✓
Digitization Type Differentiation	Classify manuscript collection images as digitized from original or microform	✓		✓	✓	✓	✓
Second-Round Explorations							
Document Clustering	Extract high-level features, cluster, investigate similarity	✓		✓	✓	✓	✓
Figure/Graph Extraction	Find and localize image-like components in newspaper pages	✓	✓		✓	✓	
Advanced Document Image Quality Assessment	Analyze quality of manuscript collection page images for compactness	✓		✓	✓	✓	✓
Digitization Type Differentiation	Fine-tune classification of manuscript collection images as digitized from original or microform	✓		✓	✓	✓	✓

6.1 Exploration: Document Segmentation

The goal of this exploration was to see if we could localize textual zones, figures, layout borders, and tables and then identify image-like components in historic newspaper pages. Currently, newspaper page images presented through Chronicling America are not zoned or segmented below the page level. In addition, content within a newspaper page is not identified or classified by genre, type, or other features. This exploration, then, was guided by the questions: how might we use image zoning and segmentation to generate additional information about newspaper pages in the Chronicling America corpus? Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers? How might machine learning projects draw on ground truth or benchmark data already generated through crowdsourcing efforts?

This exploration applied the dhSegment tool for historical document image processing to historical newspapers.[4] ResNet, a feature extractor in dhSegment, is capable of encoding an image down to a set of high-level visual features effectively and efficiently. We applied dhSegment to two sets of newspaper images, one set from the Library of Congress's Beyond Words project (http://beyondwords.labs.loc.gov/#/), which is based on Chronicling America newspapers, and one from the Europeana Newspapers Project (https://www.primaresearch.org/datasets/ENP).

In the Beyond Words project, members of the public drew rectangular zones around illustrations, photographs, comics, and cartoons in World War I-era newspapers, and users also transcribed captions for this content. In a subsequent stage of work, users could apply a typology to the graphical content, choosing among editorial cartoon, comics/cartoon, illustration, photograph, or map. Our expectation was that Beyond Words would provide ground truth data against which we might verify machine learning-based approaches to the same challenges (graphical content zoning).

We obtained a subset of 1,532 newspaper page images from Beyond Words and corresponding data for graphical-content zones. We used 1,226 images for training and 306 images for evaluation. In two different test scenarios (BW_1500_v1 and BW_1500_v2), when trained and evaluated on Beyond Words newspaper pages, we achieved best accuracy scores of 87% and 88%. Table 2 outlines results. Unfortunately, the relatively high accuracy scores are misleading upon further examination, since the model's behavior of predicting most pixels to be background pixels is guaranteed to achieve high accuracy. The low values for the best mean intersection over union (mIoU) scores verify this problematic behavior in the model, as we observe only a 26% and 24% overlap between the target class and the model's prediction.

For further exploration of the approach, we also trained and evaluated the model on a set of 481 pages from the Europeana Newspapers corpus. These newspaper page images are already zoned and segmented, with the segments classed as background, text, figure, separator or table. These classes are different from the classes in the Beyond Words dataset, which were all classes of graphical content or background. When we trained and evaluated the model on the Europeana Newspapers, we were verifying against the respective classes of the set.

In deploying on the Europeana Newspapers dataset in four scenarios (ENP_500_v1, ENP_500_v2, ENP_500_v3, ENP_500_v4), we achieved best accuracy scores of 88%, 89%, 91%, and 91%, and best mIoU scores of 64%, 64%, 69%, and 69%. In these scenarios, text regions are included in the ground-truth, and thus the model's simple guessing that everything is background is penalized. The high accuracy scores are more trustworthy in these scenarios, as further corroborated by the higher scores for mIoU.

[4] Seguin and Ares Oliveira, dhSegment; Ares Oliveira, Seguin, and Kaplan, "DhSegment."

TABLE 2. RESULTS OF PAGE SEGMENTATION WHEN TRAINING AND EVALUATING THE DHSEGMENT MODEL ON TWO SETS OF HISTORICAL NEWSPAPERS, A ROUGHLY 1500-PAGE SET FROM BEYOND WORDS/CHRONICLING AMERICA AND AN APPROXIMATELY 500-PAGE SET FROM THE EUROPEANA NEWSPAPERS COLLECTION.

Model	Train/Eval Size	Classes	Weighted Training	Pre-processing (Normalization)	Best Score	
					Accuracy	mIoU
BW_1500_v1	1226/306	0: Background 1: Editorial cartoon 2: Comics/cartoon 3: Illustration 4: Photograph 5: Map	No	No	0.87	0.24
BW_1500_v2			Yes [10; 22; 20; 18; 8; 22]		0.88	0.26
ENP_500_v1	385/96	0: Background 1: Text 2: Figure 3: Separator 4: Table	Yes [5; 10; 40; 10; 35]	No	0.88	0.64
ENP_500_v2				Yes	0.89	0.64
ENP_500_v3			No	No	0.91	0.69
ENP_500_v4				Yes	0.91	0.69

We did not conduct a broad deployment of the Europeana Newspapers model on Beyond Words/Chronicling America pages, because we did not have verifiable, commensurate ground truth across the sets. We did, however, conduct a limited test of the Europeana Newspapers-trained model (specifically ENP_500_v3) on Chronicling America page images, and the visual results are encouraging. See Figure 2, Figure 3, Figure 4, and Figure 5 for true-positive correlations, according to visual inspection, as well as for examples of false-positives and false-negatives, respectively.

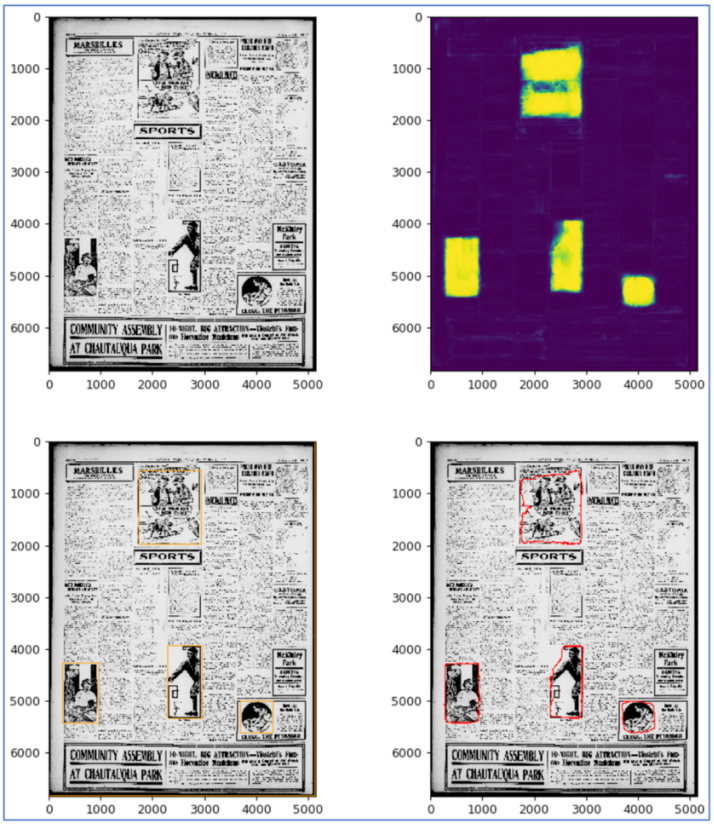


FIGURE 2. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN92053240-19190805.JPG). CLOCKWISE FROM TOP- LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

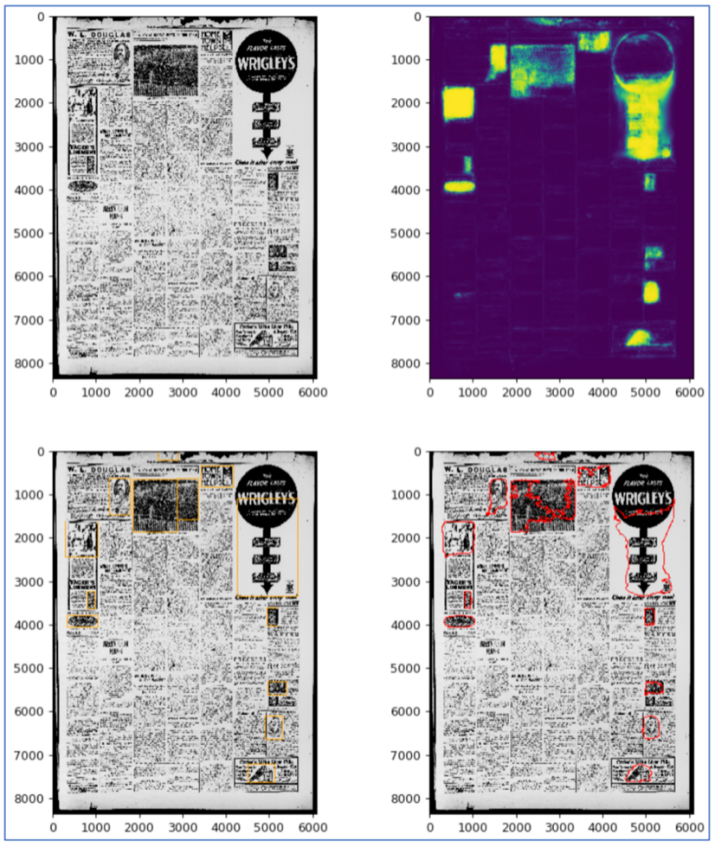


FIGURE 3. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN84026820_00271765095_1917050501_0153.JPG). CLOCKWISE FROM TOP-LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

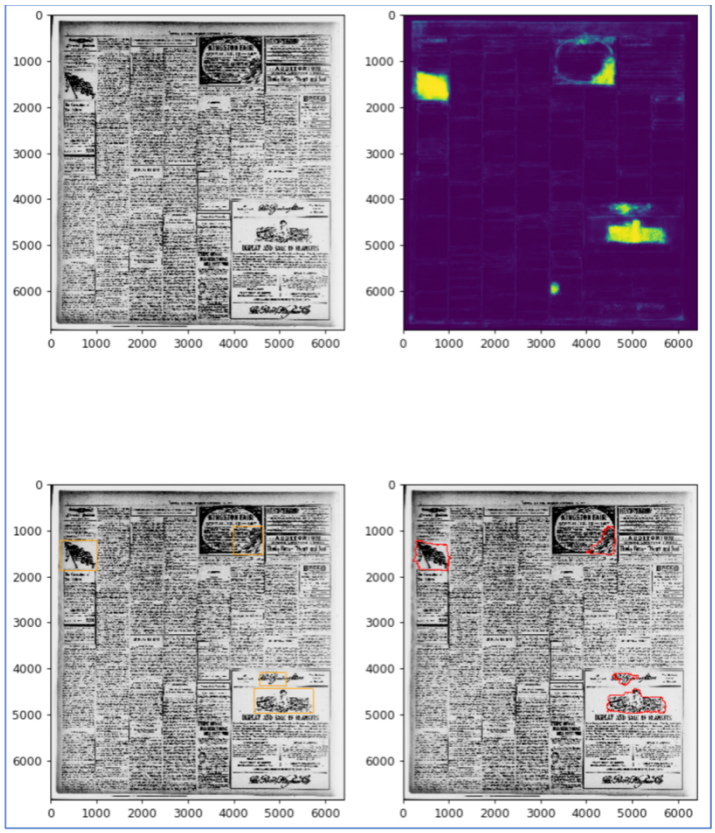


FIGURE 4. SEGMENTATION RESULT OF ENP_500_V4 ON A CHRONICLING AMERICA IMAGE (SN82014086_00295866135_1917091301_0116.JPG). CLOCKWISE FROM TOP-LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

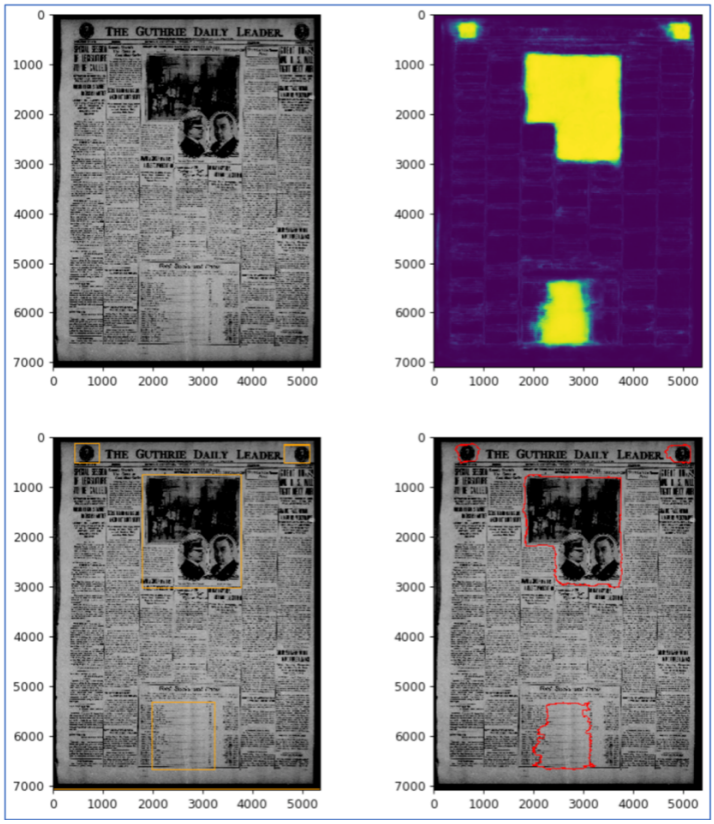


FIGURE 5. SEGMENTATION RESULT OF ENP_500_V4 ON CHRONICLING AMERICA IMAGE (SN86063952-19190805.JPG). CLOCKWISE FROM TOP- LEFT: (1) INPUT, (2) PROBABILITY MAP FOR FIGURE CLASS, (3) DETECTED FIGURES IN POLYGON, AND (4) DETECTED FIGURES IN BOUNDING-BOX. IN THE PROBABILITY MAP, PIXELS WITH A HIGHER PROBABILITY OF BELONGING TO THE FIGURE CLASS ARE SHOWN WITH A BRIGHTER COLOR.

6.2 Exploration: Graphic Element Classification and Text Extraction

This exploration was similar to the document segmentation exploration and became closer in goal and scope to that exploration over its iterations. Initially, the goal of this exploration was to find and localize figures, illustrations, and cartoons present in historical newspaper page images; classify the graphical content; and extract any text from the graphical content in order to generate a transcription of the textual content. By its second iteration, this exploration focused on fine-tuning the identification of graphical content in historic newspaper page images and the distinction of graphical content regions from textual content regions. The questions that guided this exploration throughout its development included: how might we use image zoning and segmentation, and text extraction from graphical regions, to generate additional information about newspaper pages in the Chronicling America corpus? Could image zoning and segmentation be used to pull out graphical content from Chronicling America newspapers? What benefits do different types or approaches to zoning and segmentation have for various information tasks? What strategies might be necessary to deal with rare content types in the training and evaluation of machine learning systems?

This exploration proceeded in two phases. The first phase established a conceptual model and workflow for a two-stepped approach that would result in segmented and classified graphical regions from historic newspaper pages and the segmentation and recognition of textual content in the graphical regions. Conceptually, our approach was

based on dhSegment, but instead of combining U-Net[5] and ResNet-50[6] as we did in the document segmentation exploration, we used the ResNeXt[7] classification model, a next-generation model from the ResNet that was employed in dhSegment. Such network combination belongs to the family of fully convolutional neural networks (FCN). We call our FCN that uses ResNeXt "U-NeXt." The goal was to see whether we could further enhance the results obtained with dhSegment by using this newer method. The model training was based on the pre-trained ResNeXt model for ImageNet. Finally, in the conceptual workflow, we planned to use EAST[8] text detection to find textual regions in the graphical images and use an optical character recognition process to recognize the textual strings within the graphical zones.

Our goal was to apply this conceptual model to newspaper page images from the Beyond Words project. In the above document segmentation exploration, the mIoU score was only 24%-26% on the Beyond Words dataset. We considered possible reasons that for the low mIoU scores. One possibility was that the feature extractor, ResNet, was not powerful enough to extract high-level features from the dataset for identification and classification. Notably, the ResNet model was reported by He et al. in 2015. However, in 2017, they reported a second-generation, ResNeXt, which beat the previous record on an ImageNet challenge.[9] Another possibility was that the rareness of some types of regions, such as maps, which comprise 1% of the regions, might skew the training process. As a result, we decided to test the ResNeXt model.

In addition, the data from Beyond Words were not sufficiently reliable for training purposes for this exploration.[10] One challenge with the data is that it does not include graphical content in advertisements; our model does not distinguish between graphical content in advertisements and graphical content in other types of content zones—graphical content is graphical content, at the stage of graphical content recognition and segmenting. In addition, not all graphical content has necessarily been marked on a page in the Beyond Words dataset. Since machine learning models will try to find all graphical content within the input page, such missing graphical regions can confuse the model during the training process. Another challenge, which we explore more fully below, is that ground-truth regions are not necessarily tightly mapped to the actual shape of the graphical region.

In its second phase of development, then, this exploration become a refinement of the original document segmentation exploration. We deconstructed the conceptual model described above and focused only on implementing the U-NeXt fully convolutional neural network for the purposes of graphical content extraction and classification. Our goal in doing so was to see if we could further improve upon the results reported by the dhSegment authors and the results we achieved in our document segmentation exploration that implemented dhSegment.

In the second phase of this exploration, we first conducted a pre-training investigation, which involved training and testing on the Europeana Newspapers dataset, since it is more comprehensively labeled than the Beyond Words

[5] Ronneberger, Fischer, and Brox, "U-Net."
[6] He et al., "Deep Residual Learning for Image Recognition."
[7] Xie et al., "Aggregated Residual Transformations for Deep Neural Networks." Compared to ResNet, ResNeXt uses grouped convolution (i.e. side-by-side convolution layers) in each residual block. The usage of grouped convolution was first mentioned in AlexNet. See Krizhevsky, Sutskever, and Hinton, "ImageNet Classification with Deep Convolutional Neural Networks."
[8] Zhou et al., "EAST."
[9] Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge."
[10] For more on Beyond Words, see the Library of Congress Labs' Experiments page, https://labs.loc.gov/experiments/?st=gallery Note as well that Beyond Words was not implemented with the purpose of creating training data or being used as training data.

data for our purposes. This pre-training investigation reached 91.30% pixel-wise accuracy and 57.19% for mIoU, with a testing performance of 81.90% pixel-wise accuracy and 48.18% mIoU. Note that the investigation did not reach the score of dhSegment on the ENP dataset in the document segmentation exploration. However, considering the ENP dataset and the Beyond Word dataset only share partial features, so it is not necessary to train the investigation to its best state. In fact, the observed convergence indicated the parameters were getting trained to fit the task and the model was ready for fine-tuning.

Then, we conducted a series of fine-tuning investigations, which involved four different approaches:

1. The first approach trained and tested U-NeXT on the Beyond Words dataset without using the Europeana Newspapers-trained classifier. This approach was meant to serve as a baseline design. We observed convergence in both training and testing curves, but the testing curve showed instability with rapid high and low variation during the investigation. Statistics showed that the classifier failed to recognize classes of editorial cartoons, illustrations, and maps. These three classes were the three rarest classes in the ground truth set, and the misrecognition issue is likely caused by the rareness of corresponding classes.
2. The second approach used the Europeana Newspapers-trained classifier as the beginning classifier. We then trained and tested it on the Beyond Words dataset. We added this design because using a pre-trained classifier for a similar task could help the overall fine-tuning investigations address challenges with reliable ground truth when working only with the Beyond Words dataset. Though performance indicators appeared promising, upon further investigation, the classifier trained during the fine-tuning experiment attempted to classify many pixels as background pixels after training convergence. Therefore, while the performance statistics are better than the first fine-tuning experiment numerically, the actual performance is worse, since none of the object classes (specific types of graphical content) were recognized.
3. The third approach replaced a deconvolutional layer with a resizing layer in the deep learning model. For this approach, we trained and tested on the Beyond Words dataset. This approach is designed to address a problem with the deconvolutional layer[11]; the resizing layer is perceived as an improvement on the overall technique. The pixel-wise testing accuracy is higher than in fine-tuning approach #1, but the mIoU is lower than in that fine-tuning approach. As with fine-tuning approach #2, we also found that pixel-wise accuracy and mIoU of the editorial cartoon, illustration, and map classes are zeros. However, the testing curve did not show the same instability as in fine-tuning approach #1. This result suggests that the resizing layer helped to address the challenge with the deconvolutional layer, so it is more stable than fine-tuning approach #1, though less accurate overall.
4. The fourth approach performed a two-class segmentation and classification, instead of six-class processes on the Beyond Words dataset for both training and testing. We reduced the number of classes to two because the training dataset is biased where there is a predominantly large number of background pixels compared to other classes of pixels.[12] Pixels in non-background classes comprise only 11.79% of the entire training dataset in total. By collapsing all the object pixels into one class, we can reduce the imbalance in the number of pixels in each class during training. The results indicate that training a classifier to learn information from rare classes is very hard. Combining five non-background classes into one class decreases the complexity of the task. The combined class segmentation outperformed the other fine-tuning experiments.

[11] Odena, Dumoulin, and Olah, "Deconvolution and Checkerboard Artifacts."

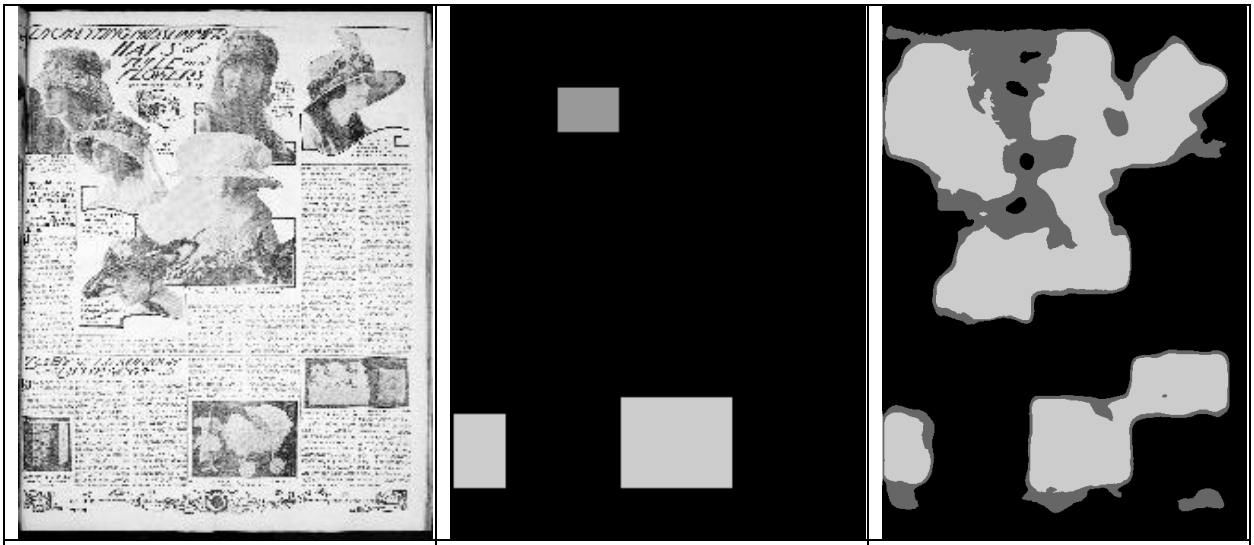
[12] There are 88.21% pixels in background class, but for the rest of classes, only 0.71% in editorial cartoon class, 2.89% in comics/cartoon class, 1.38% in illustration class, 6.64% in photograph class, and 0.18% in map class.

Approaches two through four are variants of the first approach. See Table 3 for a summary of results for the pre-training investigation and the four fine-tuning investigations.

TABLE 3. AVERAGE PERFORMANCE OF PRE-TRAINING INVESTIGATION AND FINE-TUNING APPROACHES.

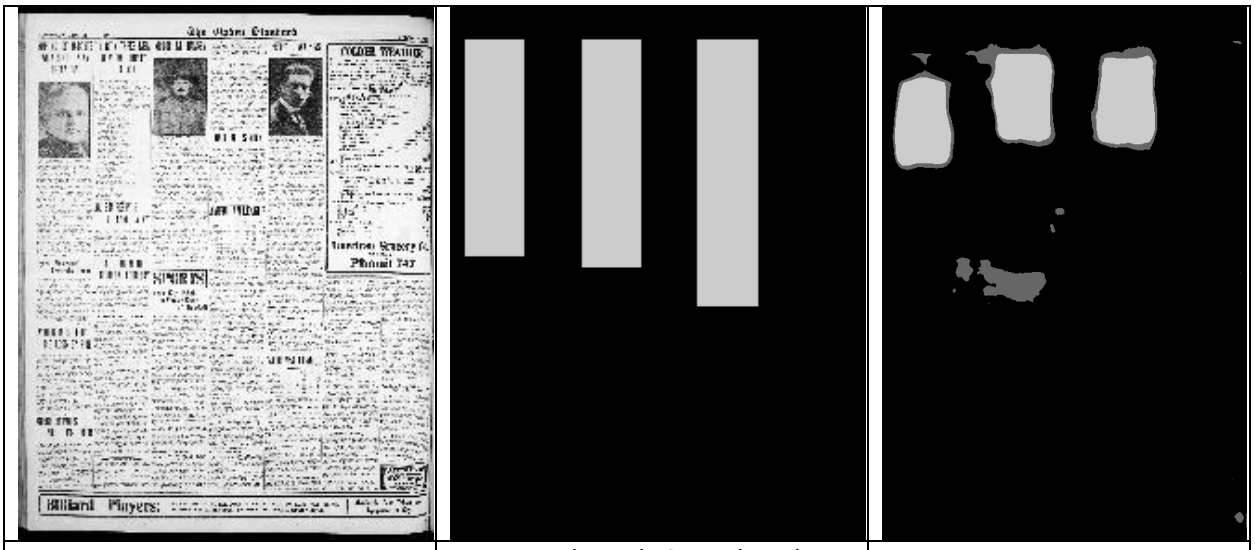
	Pre-Training Investigation		Without Pre-Trained Europeana Newspapers Classifier		Using Pre-Trained Europeana Newspapers Classifier		Using Resizing Layers		Combined Two-Class Segmentation	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Accuracy	91.30%	81.90%	89.08%	80.11%	89.42%	85.53%	88.90%	86.69%	91.76%	88.89%
mIoU	57.19%	48.18%	50.43%	38.00%	41.21%	38.57%	51.31%	37.84%	71.44%	64.97%

From these investigations, we conclude that U-NeXt—especially the combined two-class segmentation—is promising for segmentation and zoning. At the same time, the fine-tuning approaches offered evidence that the Beyond Words dataset was not sufficient ground truth for our purposes. We found two issues. First, non-identified or incompletely identified graphical images in the Beyond Words dataset appear to be widespread to an extent that is problematic for training. For example, as shown in Figure 6, a large portion of a photograph in the document is missing from the ground truth but is captured by our U-NeXt classifier. Second, the rectangular regions do not necessarily match to the actual graphical content. For instance, as shown in Figure 7, the ground truth region includes a large portion of the textual content.



Newspaper Page | Beyond Words Ground Truth | U-NeXt Map

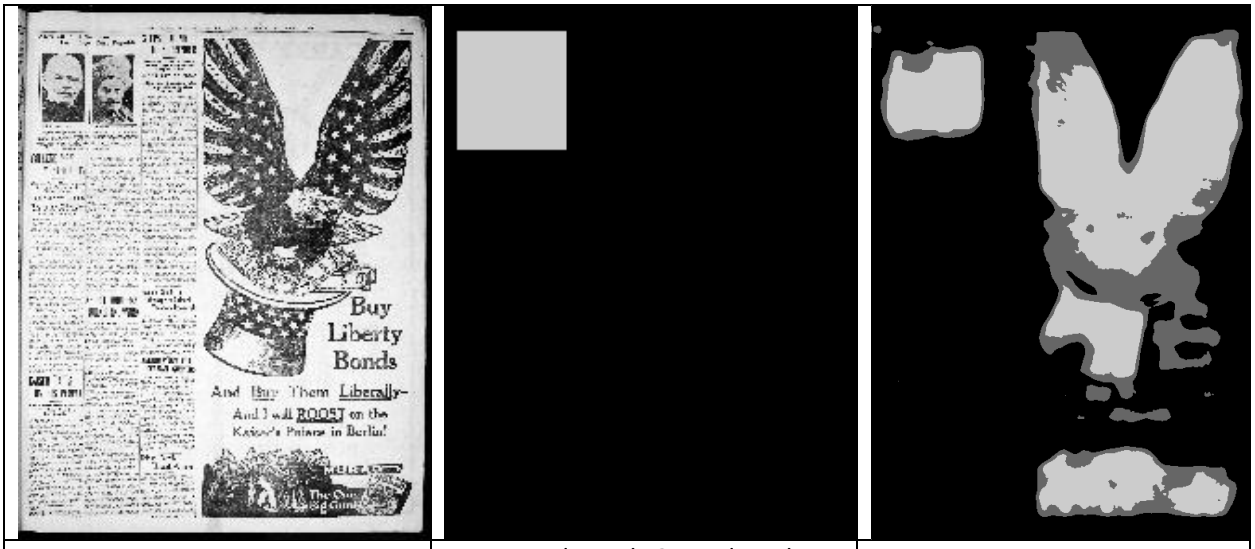
FIGURE 6. THE BEYOND WORDS DATA, WHICH WE TREATED AS GROUND TRUTH, IS MISSING MUCH OF THE GRAPHICAL CONTENT ON THE PAGE, WHILE THE U-NEXT MAP APPEARS MORE REPRESENTATIVE OF THE ORIGINAL PAGE.



Newspaper Page	Beyond Words Ground Truth	U-NeXt Map

FIGURE 7. EACH OF THE THREE COLUMNS WITH GRAPHICAL CONTENT HAVE MUCH LARGER BOUNDING BOXES IN THE GROUND TRUTH THAN WHAT CORRESPONDS TO THE ACTUAL GRAPHICAL CONTENT. THE U-NEXT MAP APPEARS MORE REPRESENTATIVE OF THE ORIGINAL.

These challenges with the Beyond Words dataset as ground truth for this exploration also lead us to believe that our classifier may be more accurate than the current statistical results would suggest. The U-NeXt model tries to fit the exact shape of the figure and graph region. Figure 8 shows, for example, that the model tried to fit the exact shape of the eagle on the right-hand side of the newspaper page. Since the ground truth included rectangular bounding boxes, we are not comparing like to like in our pixel-wise and mIoU comparisons.



Newspaper Page	Beyond Words Ground Truth	U-NeXt Map

FIGURE 8. IN THIS CASE, THE "BUY LIBERTY BONDS" ADVERTISEMENT IS NOT REPRESENTED IN THE GROUND TRUTH, LIKELY BECAUSE IT IS AN ADVERTISEMENT AND THEREFORE OUT OF SCOPE FOR THE BEYOND WORDS PROJECT. THE U-NEXT MAP, HOWEVER, RECOGNIZES THE GRAPHICAL CONTENT AND CLOSELY FITS TO IT.

6.3 EXPLORATION: DOCUMENT TYPE CLASSIFICATION

This exploration pursued whether we could effectively distinguish among handwritten, printed, and mixed (both handwritten and printed) documents within a collection of minimally processed manuscript materials at the Library of Congress. This exploration was guided by the questions: what features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials? How viable might large-scale indexing of documents be, for certain types of criteria? To what level of performance could we meta-tag document images? Would a deep learning model that had shown remarkable performance for natural scene images also show promising performance for document images? Or, to be more precise, would a feature extractor trained with millions of natural scene images also capably extract useful features for document images?

This exploration drew on current state-of-the-art methods in natural image and document classification. In particular, we extended the use of convolutional neural networks for classifying natural images to the task of classifying document images. Based on the findings of Harley et al. and Afzal et al., we used the VGG method with 16 categories (VGG-16) pre-trained on the Ryerson Vision Lab Complex Document Information Processing (RVL_CDIP) dataset.[13] The *RVL_CDIP* dataset, which is publicly available, consists of 400,000 document images that are divided into 16 evenly distributed classes. The dataset is provided in three different sets: training, validation, and test set. The training set contains 320,000 images of 16 different evenly distributed classes (i.e., about 20,000 images per class). Both validation and test sets together contain 40,000 images of 16 different evenly distributed classes (2,500 images per class).

We first set out to reproduce the results reported in the work of Harley et al. and assessed classification performances of VGG-16, pre-trained on ImageNet, and trained and tested with *RVL_CDIP* dataset. The advantage of doing so is that once we created a model trained on this large-scale document image dataset, we can reuse the rich features that this model has learned for many document analysis tasks, such as for our current ask of document type classification. The entire training process took only three epochs to converge with promising classification results. This indicates that features obtained from natural scene images (i.e., ImageNet) are general enough to be applied to documents. The resultant classification performance metrics—precision, recall, and f1-score—are shown in Table 4. On average, each metric shows around 87%, which aligns well with the result reported by Harley et al.

TABLE 4. PRECISION, RECALL, AND F1-SCORE OF VGG-16 AS TRAINED ON RVL_CDIP DATASET. THE ALPHABETIC LABELS CORRESPOND TO THE FOLLOWING LABELS: LETTER, FORM, EMAIL, HANDWRITTEN, ADVERTISEMENT, SCIENTIFIC REPORT, SCIENTIFIC PUBLICATION, SPECIFICATION, FILE FOLDER, *NEWS ARTICLE*, *BUDGET*, *INVOICE*, *PRESENTATION*, *QUESTIONNAIRE*, *RESUME*, AND *MEMO*. OUR CLASS OF INTEREST, HANDWRITTEN, IS BOLDED.

	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Avg
Precision	86	74	98	**89**	89	73	90	99	89	92	87	91	78	91	92	88	87
Recall	94	79	97	**96**	91	73	93	91	97	86	83	86	79	73	94	91	87
F1	86	77	97	**92**	90	73	91	90	93	89	85	88	79	81	93	90	87

Next, we generated our own model for the specific task of classifying documents in one of three types, handwritten, typed/typeset, or mixed (both handwritten and typed/typeset). For this task, we retrained the model obtained from

[13] Harley, Ufkes, and Derpanis, "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval"; Afzal et al., "Cutting the Error by Half"; Simonyan and Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition."

the above with a dataset derived from the Suffrage: Women Fight for the Vote campaign from By the People. The dataset for this exploration is hereafter referred to as suffrage_1002.

The suffrage_1002 dataset had 1,002 manually classified images and was balanced across handwritten, typed/typeset, and mixed. This ground truth set was created by members of the project team. The entire dataset was split into three sets—training, validation, and test—with a ratio of 8:1:1. In order to keep the class balanced during this split, we dropped three datapoints, one of each class. The final size of the dataset was therefore 999 images. See Table 5 for the breakdown by sets and class.

TABLE 5. CONFIGURATION OF SUFFRAGE_1002 DATASET.

	Handwritten	Typed/Typescript	Mixed	Total
Train	267	267	267	801
Validation	33	33	33	99
Test	33	33	33	99
Total	333	333	333	999

We use the same *VGG-16* architecture as above, but the output tensor was adjusted to have a shape of 3, the number of classes specified in suffrage_1002. Overall, our model's classification performance on the testing set shows about 90% for precision, recall, and f1-score, as shown in Table 6. We believe that these scores, which are a bit lower than those reported above in our attempts to reproduce Harley et al., are due to challenging characteristics of *mixed* type document images; for example, mixed materials may have negligible or statistically challenging amounts of handwriting in typed document and vice versa. See Figure 9 for examples.

TABLE 6. PRECISION, RECALL, AND F1-SCORES OF VGG-16 ON SUFFRAGE_1002 TESTING SET.

	Handwritten	Typed/Typescript	Mixed	Avg
Precision	89	91	90	90
Recall	97	94	79	90
F1	93	93	84	90

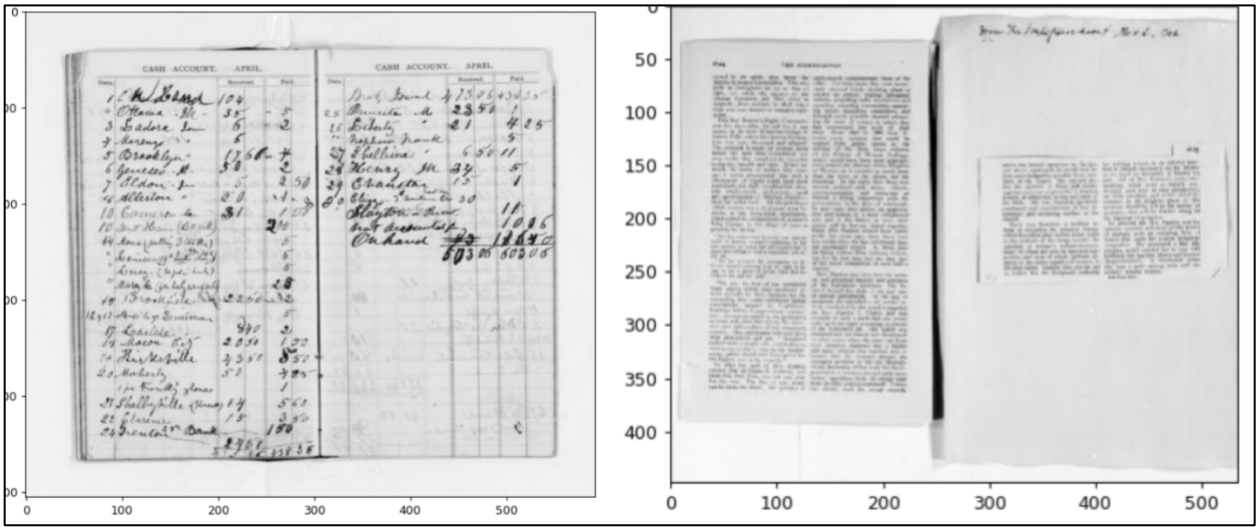


FIGURE 9. PREDICTION FAILURE CASES. IN THE LEFT EXAMPLE, THE MODEL CLASSIFIED THE DOCUMENT AS HANDWRITTEN RATHER THAN MIXED. NOTE THAT THE PRINTED REGIONS ARE VERY SMALL COMPARED TO THE HANDWRITTEN CONTENT IN THE IMAGE. IN THE RIGHT

EXAMPLE, THE MODEL CLASSIFIED THE DOCUMENT AS PRINTED RATHER THAN MIXED. HERE, THE HANDWRITTEN REGION IS VERY SMALL COMPARED TO THE PRINTED REGION IN THE IMAGE.

In the case of the examples in Figure 9, both images have technically been classified incorrectly, according to our current model and set of definitions. Both images depict documents that feature printed and handwritten content, and therefore both technically fit the definition of mixed. However, this example provides an opportunity to consider whether the actual mixed nature of these materials matters for processing purposes. If the Library of Congress were interested in this question for the benefit of helping to make decisions about how to handle particular types of materials—for example, that materials with significant handwritten content get passed to human experts, whether within the Library or outside of it—then materials with some, but limited print content are usefully grouped with handwritten materials. Likewise, if there was a strong mix of content types, that might also signal materials for human processing, whereas in the example of materials with minimal handwritten content, it may be fine to pass those materials off to more automated processes.

6.4 Exploration: Digitization Type Differentiation

The purpose of this exploration was to distinguish among digital images created through digitization from different source types. In particular, we sought to distinguish between items digitized from an original document item and those digitized from a microform reproduction of an original item. We expected that digitization source should be a relatively easy feature to distinguish and could have a variety of potential use cases for both internal processes and decision-making at the Library and for end users and researchers. A variety of questions sat behind this exploration. As with the document classification exploration, we wondered: what features might be useful for influencing processing pipelines, for generating additional metadata, or for distinguishing among materials? How viable might large-scale indexing of documents be, for certain types of criteria? To what level of performance could we meta-tag document images? We also wondered who might benefit from the ability to facet or search according to this particular criterion—digitization source—and how that might information might be made available.

This exploration proceeded in two phases. In both, we used ResNeXt, a deep learning method, to differentiate among images digitized from an original and those digitized from a microform reproduction. All images for this exploration came from the minimally processed manuscript collections included in the By the People Civil War campaign.

We first retrieved 36,103 images from the minimally processed Civil War materials and manually inspected 10,508 images, or slightly less than 30% of the total images. We determined digitization source ground truth for each of these 10,508 images. We then sampled a subset of 1,200 images from the 10,508 in a balanced set (600 images of each type, digitized from original and digitized from microform). In a 10% test of 120 sample images, the classifier was 100% accurate in classifying images as digitized from an original item or from a microform. We had concern, however, that this 100% accuracy was likely too good to be true when deployed over a larger set of images. We therefore proposed to compare the ratio of items digitized from microform to items digitized from original items to more comprehensively evaluate our approach. Based on the ground truth classification of the 10,508 images, we would expect a 1:16 ratio of images digitized from microform to images digitized from original items across the Civil War dataset.

In the second phase of this exploration, we fine-tuned the classifier, classified 36,103 images retrieved from the Civil War manuscript collections, and compared the number and ratio of expected classification to real classifications. In fine-tuning the classifier, we achieved a training accuracy of 98.52%, and a validation accuracy of 100%. In order to determine an ideal point between underfitting and overfitting the classifier, we calculated the harmonic mean of training performance and validation performance, to avoid both underfitting and overfitting.

We then used the fine-tuned classifier to classify the 36,103 images. Based on the first phase of this exploration, we expected that the image set would include 2,256 document images digitized from microfilm and 33,847 digitized from their original source. In reality, the classifier identified 2,834 images as digitized from microform and 33,269 images as digitized from an original item. See Table 7. Therefore, while we expected a classification ratio of 1:16 (microform to original), the achieved classification ratio was 1:11.74. The classifier was more aggressive in identifying images as having been digitized from a microform reproduction than we would have anticipated based on our initial tests.

TABLE 7. BREAKDOWN OF PROJECTIONS AND ACTUAL CLASSIFICATIONS OF CONTENT AS DIGITIZED FROM MICROFORM OR DIGITIZED FROM AN ORIGINAL ITEM.

Total Images	Expected Microform Source	Classified Microform Source	Expected Original Source	Classified Original Source
36,103	2,256	2,834	33,847	33,269

Without identifying the ground truth of each of the 36,103 items, we cannot be sure if the 1:16 ratio is entirely accurate. However, we do know that the classifier was more aggressive in identifying items as digitized from microform reproductions than in classifying them as digitized from an original item. For example, each of the items in Figure 10, Figure 11, Figure 12, and Figure 13, while actually digitized from original items, were classified as being digitized from microform reproductions.

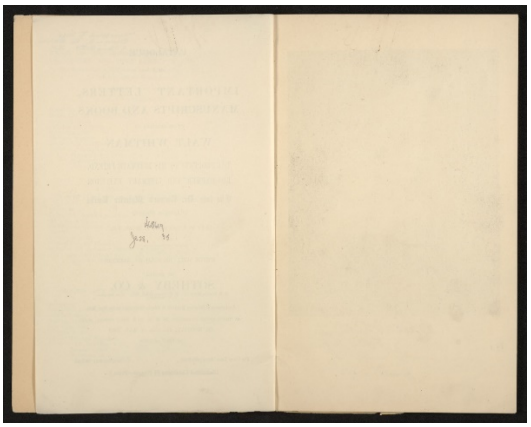


FIGURE 10. FACING PAGES OF A DOCUMENT, DIGITIZED FROM THE ORIGINAL, THAT THE CLASSIFIER CLASSIFIED AS HAVING BEEN DIGITIZED FROM MICROFORM. THERE IS MINIMAL CONTENT ON THE PAGES.

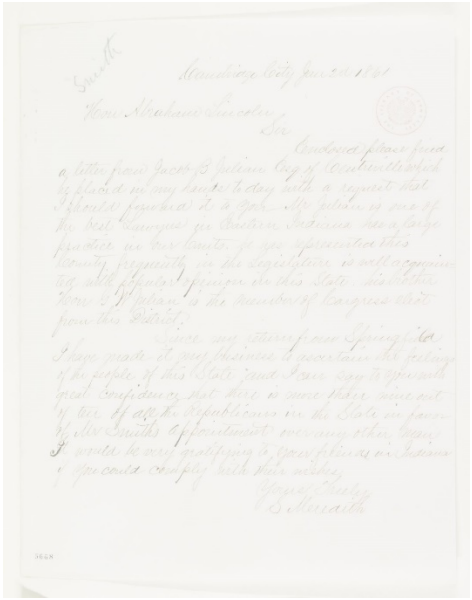


FIGURE 11. A HANDWRITTEN MANUSCRIPT PAGE IMAGE, DIGITIZED FROM THE ORIGINAL, AND CLASSIFIED BY OUR CLASSIFIER AS HAVING BEEN DIGITIZED FROM MICROFORM. THE CONTRAST OF THE IMAGE IS LOW.

FIGURE 12. A DIGITAL IMAGE OF A COIN, DIGITIZED FROM THE ORIGINAL ITEM, AND CLASSIFIED BY OUR CLASSIFIER AS HAVING BEEN DIGITIZED FROM A MICROFORM REPRODUCTION.

FIGURE 13. A DIGITAL IMAGE OF A PHOTOGRAPH, DIGITIZED FROM THE ORIGINAL ITEM, AND CLASSIFIED BY OUR CLASSIFIER HAS HAVING BEEN DIGITIZED FROM A MICROFORM SOURCE.

We believe these misclassifications were due to limitations of our training set, which did not include blank pages digitized from original items, photographs (that is, photographs of people that are included in the manuscript collections, for example), images with poor document contrast, or 3-d objects represented in the collections.

In the future, two options could effectively improve the performance further. First, we can expand the training database to include more examples and types of materials. Second, we can apply a pre-processing step to normalize the document image quality for the collection before the prediction stage. Overall, however, results were promising and suggest that automated type differentiation is viable and computationally cheap.

6.5 Exploration: Document Image Quality Assessment (DIQA) & Advanced DIQA

This exploration set out to analyze the quality of document images in minimally processed manuscript collections based on a variety of criteria with the goal of using information about image quality to inform future processes and toward making this information available for researchers looking for particular kinds of images (or images of particular quality). This exploration was guided by the questions: how might we distinguish among materials that most need human intervention, whether by Library of Congress staff or via crowdsourcing and the public, and those materials that might be well-suited to machine approaches? And when might materials be best suited to a combined approach? Could image quality assessments be useful in compiling ground truth and benchmarking sets in some capacity? Likewise, might such features be useful further downstream for users, to be able to facet for difficulty, for example? How might metadata about image quality of document images enrich understanding of individual items and of collections and corpora? To what extent can quality be computationally assessed, and might it help to better understand overall visual attributes of a dataset?

This exploration proceeded in two phases. In the first phase, we measured a set of image properties for each of 35,990 images retrieved from minimally processed manuscript collections included in the By the People Civil War

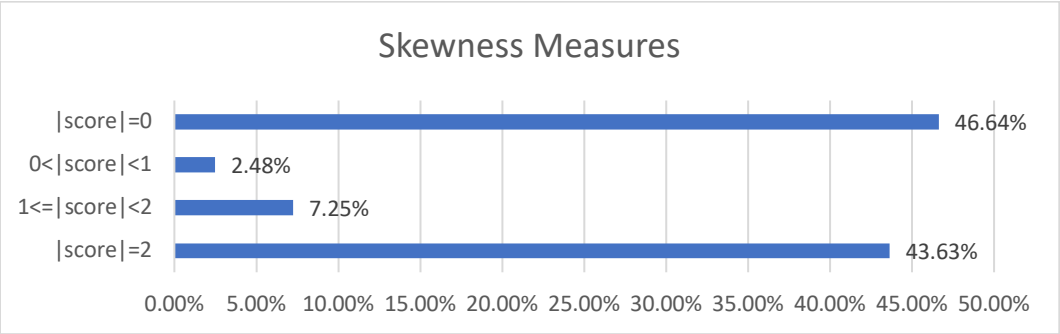
campaign. These image properties included skewness, contrast, range-effect, and bleed-through (background noise).[14]

6.5.1 Document Image Quality Assessment (DIQA)

6.5.1.1 Skewness

The skewness measure ranges from a score of -2 to 2, with any score other than 0 indicating skew is present. For example, a skew of -2 indicates significant counterclockwise skew, while a score of 2 indicates significant clockwise skew. Of the 35,990 images, nearly 50% of the images show no or negligible skew (a score of 0, or between 0 and |1|). Nearly 44% of the images (43.63%), are significantly skewed. See Chart 1.

CHART 1. SKEWNESS MEASURES OF 35,990 IMAGES FROM MINIMALLY PROCESSED CIVIL WAR COLLECTIONS. 43.63% OF IMAGES ARE SIGNIFICANTLY SKEWED (SCORE |2|).



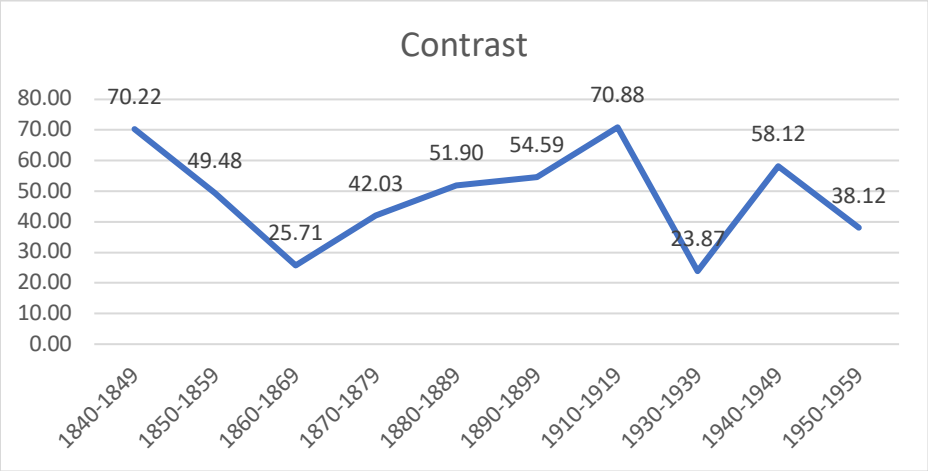
6.5.1.2 Contrast

Based on earlier work, the Aida team has determined that a contrast score of 30 or above indicates a good quality contrast in a digital image of a historic document;[15] the higher the contrast score, the better the visual quality. We plotted contrast over time (original date of document page, based on existing metadata) and determined that the two decades of materials represented in the Civil War collection fell below the threshold for a good contrast score: 1860–1869 and 1930–1939. See Chart 2.

[14] The document image quality assessment algorithms used in this exploration were developed as part of the Aida team's earlier efforts to assess qualities of newspaper page images from 1834 to 1922. See Lorang, Soh, Liu, Pack, and Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures."

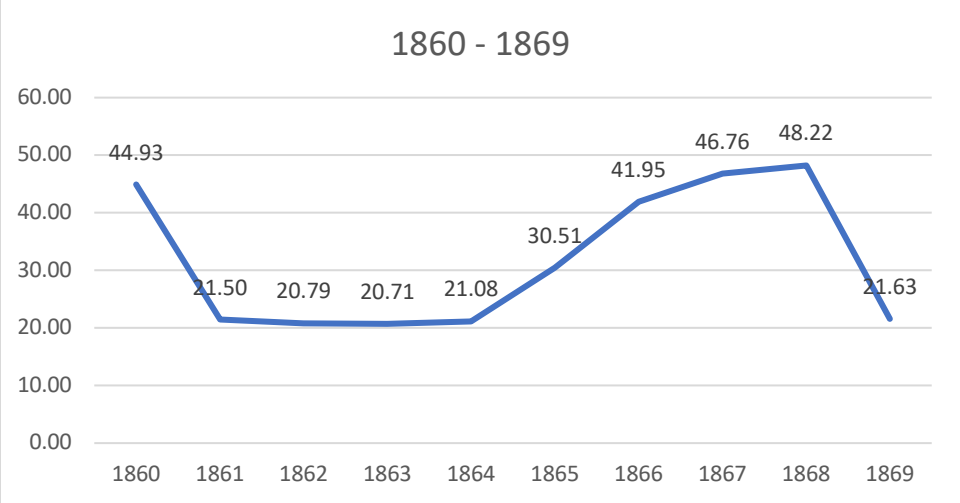
[15] Lorang, Soh, Liu, Pack, and Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures."

CHART 2. AVERAGE CONTRAST SCORES OF MATERIALS WITHIN DECADE-RANGES. FOR EXAMPLE, MATERIALS FROM THE PERIOD 1840-1849 HAVE AN AVERAGE CONTRAST SCORE OF 70.22, WHILE MATERIALS FROM THE PERIOD 1930-1939 HAVE AN AVERAGE CONTRAST SCORE OF 23.87.



While images from 1930 to 1939 result in lowest contrast score, the significant majority of images in the collection—roughly 90%—date to between 1860-1869. These dates also overlap with the actual years of the Civil War, making the images from that decade most critical for further analysis. When we look more closely at the contrast scores within this decade, 1861, 1862, 1863, and 1864 all show average contrast scores below 22. See Chart 3. These data suggest that materials from most of the actual Civil War years have the lowest contrast in the collection and also that their contrast is below the threshold for good visual contrast. The low contrast can make these materials challenging for computational processing and also for human readers.

CHART 3. AVERAGE CONTRAST SCORES OF MATERIALS FROM THE DECADE 1860-1869 BY YEAR. MATERIALS FROM 1861 HAVE AN AVERAGE CONTRAST SCORE OF 21.50, FOR EXAMPLE, WHILE MATERIALS FROM 1868 HAVE AN AVERAGE CONTRAST SCORE OF 48.22.



We suspect that the low score could be document images that are digitized from handwritten letters, shown in Figure 14. There are two persistent features among these letters that could lower the contrast score. First, the original paper is often yellowish in hue. Second, the ink or pencil is often light—again, whether due to original inscription or time, or a combination of elements.

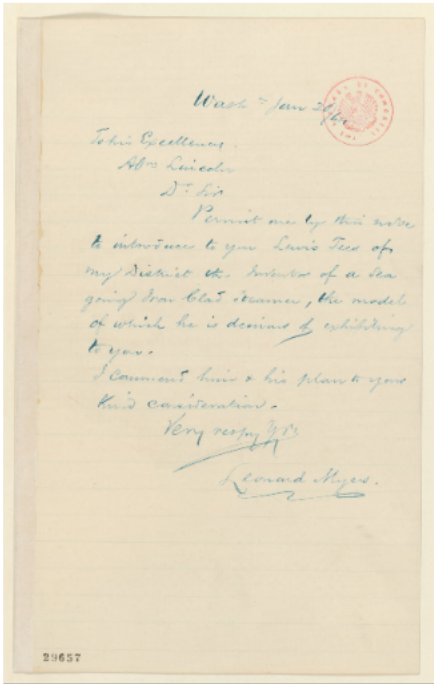
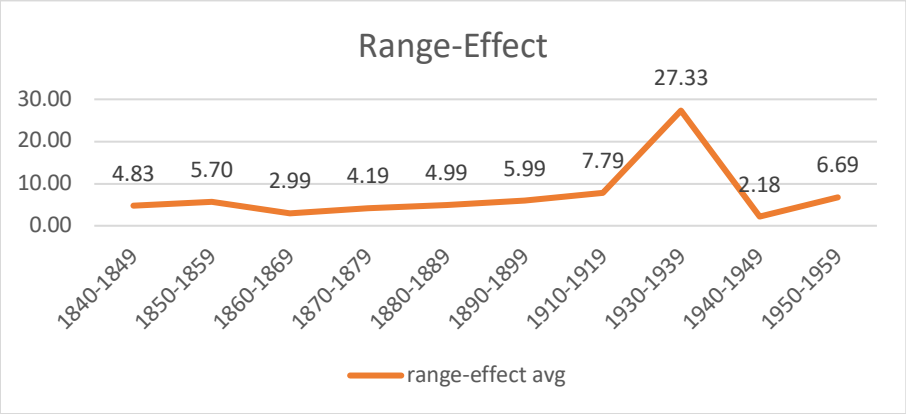


FIGURE 14. AN IMAGE WITH A LOW-CONTRAST SCORE FROM THE CIVIL WAR YEARS. FEATURES OF THIS IMAGES, SUCH AS THE COLOR OF THE PAPER AND THE LIGHT INK INSCRIPTION, MAY BE COMMON ACROSS THE MATERIALS, LEADING TO OVERALL LOW CONTRAST SCORES.

6.5.1.3 RANGE-EFFECT

The lower the range effect score, the better the quality of the image with regard to this feature. An ideal score is zero, and our team's earlier work with historic newspapers suggests that a range-effect score lower than three is indicative of a good-quality image. The materials from the decade 1860-1869 have an average range-effect score of 2.99. With materials in this decade comprising 90% of the images in the set, we believe range effect is not a substantial challenge for images in the Civil War collections. There are some noticeable outliers, as show in Chart 4, though in the case of the images from the period 1930-1939, the average score was significantly affected by two images with very high range-effect scores.

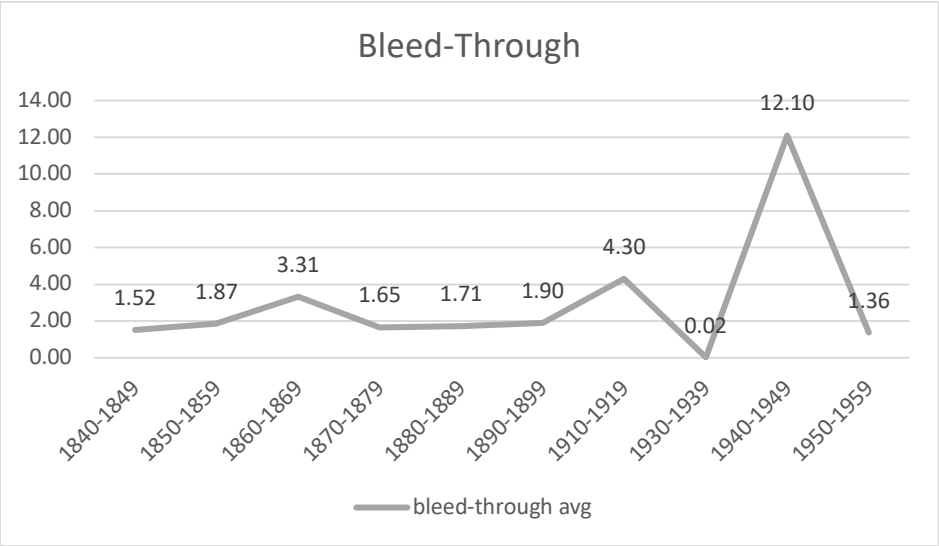
CHART 4. AVERAGE RANGE EFFECT OF THE CIVIL WAR COLLECTION OVER TIME. OVERALL, RANGE EFFECT IS LOW, WITH A SPIKE IN 1930-1939, LIKELY ATTRIBUTABLE TO A COUPLE OF DOCUMENTS IN A SMALL SET OF MATERIALS FROM THAT DECADE.



6.5.1.4 BLEED-THROUGH

Our analysis indicates that very few of the images studied suffer from significant bleed-through, which is a measure of noise in the overall image. While we do not have an objective measure for a good bleed-through score (meaning minimal bleed-through is present), an ideal score is zero. The majority of the 35,990 images return low bleed-through scores. See Chart 5. In this test, there are 76 images from the decade 1940-1949 that have high bleed-through scores and cause the average to spike in that decade.

CHART 5. AVERAGE BLEED-THROUGH/NOISE IN MATERIALS FROM THE CIVIL WAR COLLECTION, BY DECADE.



One caveat, however, is that in our processing, a document image is first converted into a grayscale image by the evaluation algorithm. Many of the pages in the collection are a yellow-hued paper that results in a dark background after the conversion. The presence of a dark background affects bleed-through evaluation and may result in a faulty evaluation.

6.5.2 ADVANCED DIQA

In the second iteration of DIQA, we combined elements of the original DIQA, document type differentiation, and document image segmentation explorations. Specifically, we measured a document image's compactness. In this case, compactness represents the number of zones (i.e., text blocks, figure) in a document image and may be considered as an indicator of document complexity.

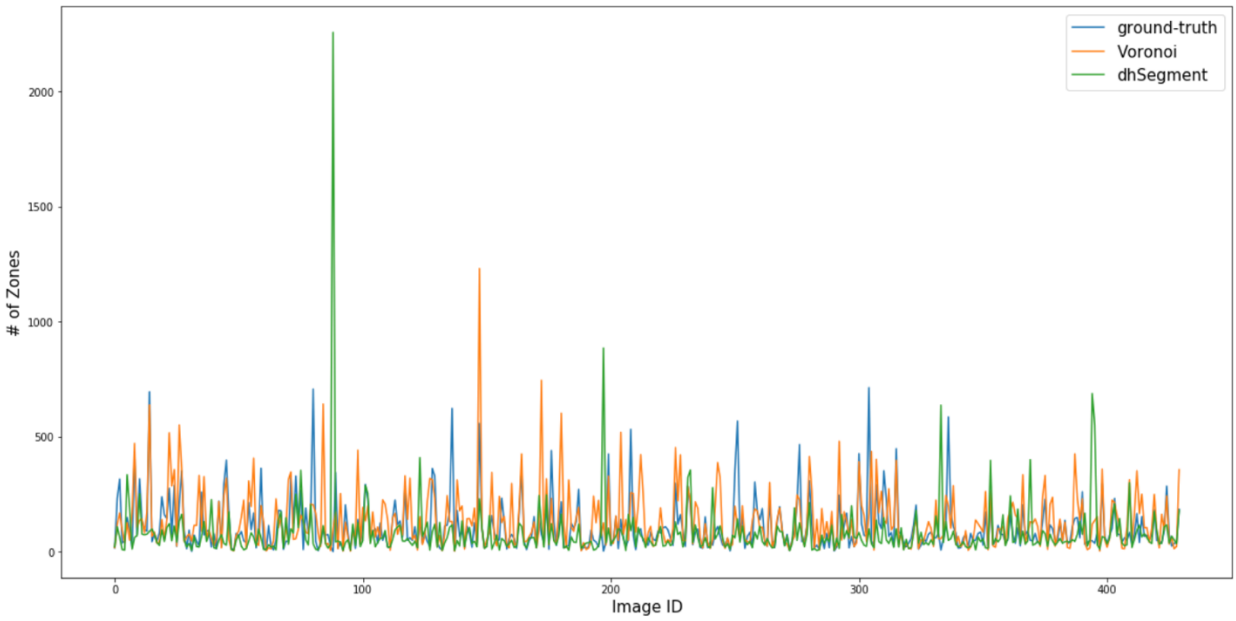
To proceed, we first assessed the compactness measures obtained by two segmentation algorithms on a dataset with known, reliable ground-truth. We then applied the compactness measure to minimally processed manuscript collections from the Civil War. Finally, we compared our measure of complexity with difficulty scores ascribed to materials through the Library of Congress's By the People site, to see if we could determine a correlation between our measure of complexity and document difficulty scores.

The two document segmentation algorithms used were (1) non-machine-learning, Voronoi diagram-based algorithm; and (2) machine-learning-based algorithm, dhSegment, used in other explorations reported here. We tested the two segmentation algorithms on 430 document images collected from the Europeana Newspapers dataset and counted the number of regions segmented by each algorithm. Then, we compared the result with the number of actual regions stored in the ground truth. See Table 8 and Chart 6.

TABLE 8. COMPARISON OF ACCURACY OF COMPACTNESS OF TWO ALGORITHMS.

	Voronoi-based Segmentation	Deep learning-based Segmentation
Mean Differences	75	77
STD Differences	97	154

CHART 6. THE COMPACTNESS OF THE EUROPEANA NEWSPAPERS DATASET.



Based on the mean and standard deviation of the difference between the number of zones detected by the algorithms and the ground truth, the compactness obtained by the non-machine-learning-based segmentation algorithm is slightly better than the machine-learning-based algorithm. In addition, when we looked at the number of zones and the compactness of the newspapers dataset, even though the number of zones detected by the two algorithms does not perfectly match with the number of actual zones, we observed a certain degree of similarity between compactness and the actual busyness of document images.

With this understanding, we then turned to several minimally processed manuscript collections from the Civil War, which are featured on By the People and are part of the Civil War campaign. Specifically, we used the Veroni-based approach to assess the compactness of four digitized document image collections, identified as “Civil War,” “Clara Barton,” “Letters to Lincoln,” and “Walt Whitman.” We analyzed within collections and by year/across time.

Three collections—Civil War, Letters to Lincoln, and Walt Whitman—show similar compactness distributions. The Clara Barton collection shows a thicker tail, which indicates that images in this collection tend to have a busier layout. Likewise, the Clara Barton collection shows notable changes in compactness across time represented in the collection. Our results suggest that items in the Clara Barton collections from 1862 to 1869 have more complex and busier layouts than those from 1850 to 1861. In the other three collections, we did not find notable compactness differences by time period.

Finally, we wondered if we might find a correlation between our compactness score and the difficulty score applied to an image through the By the People platform. We wondered: did the compactness, or busy-ness, of an image, which can be understood as a marker of complexity, correspond to images with higher difficulty scores? Ultimately, we could not correlate our measure of compactness with the difficulty score from By the People.

The difficulty score itself may not match human perception of difficulty as complex, non-linear relationships exist among visual features. In general, image quality assessment includes both machine and human perceptions of quality of an image. For machine perception, quality assessment evaluates difficulties to predict or categorize an image for a machine. And for human perception, quality assessment evaluates difficulties in understanding and interpreting an image based on the visual appearance.

6.5.3 Potential Applications

Much of this this exploration did not apply machine learning and instead was purely an image processing and image analysis exploration. We pursued it as part of this machine learning project in order grapple with which types of investigations require machine learning and when might other computational approaches be helpful to doing more with digital collections. In addition, such an exploration can facilitate future machine learning applications and endeavors. In cultural heritage digital libraries, administrative and descriptive metadata are common, even if the descriptive metadata are often limited. Various of our other explorations throughout this demonstration project, such as approaches to segmentation and classification, are toward enriching descriptive metadata and also have implications for enhanced structural metadata. As researchers begin to process large quantities of document images to develop robust classifiers or to develop generalizable automated systems, there is an increasing need for metadata about the image quality of the digitized document images, such as average intensity of an image, contrast, range effects, layout structure, etc., such that researchers might query and retrieve specific subsets of document images based on these qualities for testing.

6.6 Exploration: Document Clustering

This exploration extended from the initial documentation segmentation exploration and applied clustering to document images. Drawing on our other work with ResNet and dhSegment, we wondered whether document images clustered together share similar visual features recognizable to human observers. For example, would page images with graphical content cluster? Could we discern other clustering features? Could such clusters be useful in decision-making, for metadata generation, or other processes?

Two assumptions shaped this exploration. The first was that the deep visual representation of each datapoint contains enough feature information to be clustered. Second, in the clustered manifold, datapoints residing in the same neighborhood will share similar visual metadata with one another.

This exploration proceeded in two parts. In both parts, we used dhSegment to extract high-level visual features and then clustered the features using t-SNE, a state-of-the-art clustering method.[16] The dataset was a set of 96 page images from the Europeana Newspapers collection. From each of the 96 page images, we extracted a set of feature maps—so-called latent space—learned by a deep model, the ResNet-50 + U-Net that we trained for the document segmentation exploration. In this approach, the size of the latent space is calculated by the formula image width/32 x image height/32 x 2049.

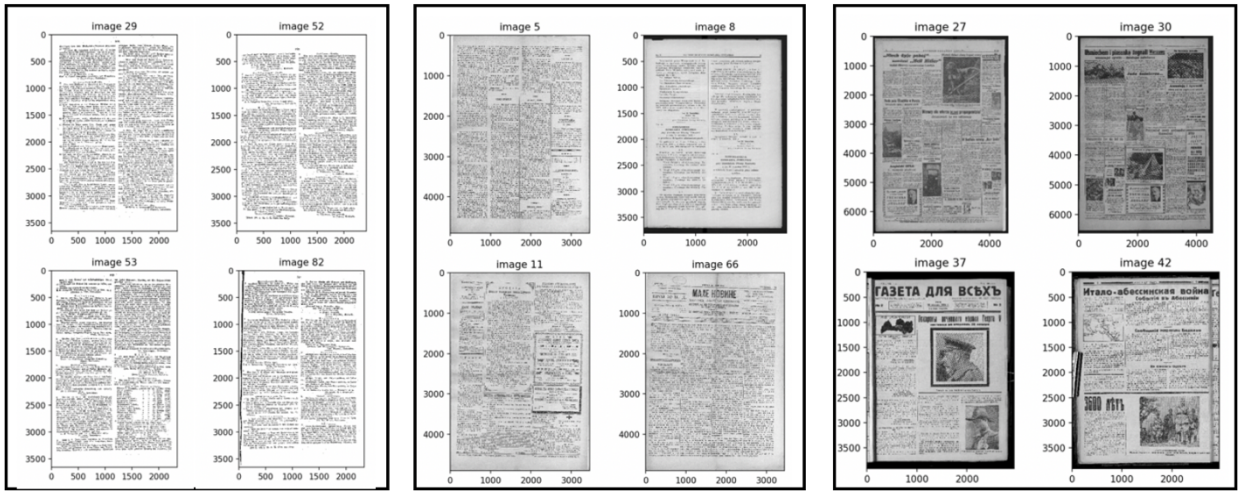
This exploration faced two challenges from the outset. First, the sizes of the input images varied, and they could not be reduced to the same proportions without distorting the images and thus their visual features. The size differences

[16] Maaten and Hinton, “Visualizing Data Using T-SNE.”

were due to variation in the original dimensions of the newspapers as well as to variations introduced during the microform process (how much non-newspaper space was captured in the duplication process, for example). These differences meant that the size of our latent space was inconsistent. A second challenge for this exploration was that the latent space was too large. The resolution of an input image might be 1800 x 2400, meaning that the corresponding latent space became 1800 x 2400 x 2408. This scenario would contain redundant information and degrade clustering performance in both quality and computation time. To address these challenges, we performed dimensionality reduction recorded the intensity of features but not their spatial location.

These reduced latent space feature maps were then clustered using t-SNE. Ninety-six (96) datapoints in 2048-dimensional space grouped into roughly three clusters in two-dimensional space. Once the images were clustered in this low-dimensional space, we visually inspected and analyzed for (1) *intraclass correlation*, or whether datapoints in the same cluster share similar visual features, and (2) *interclass correlation*, or whether different clusters show dissimilarity to each other.

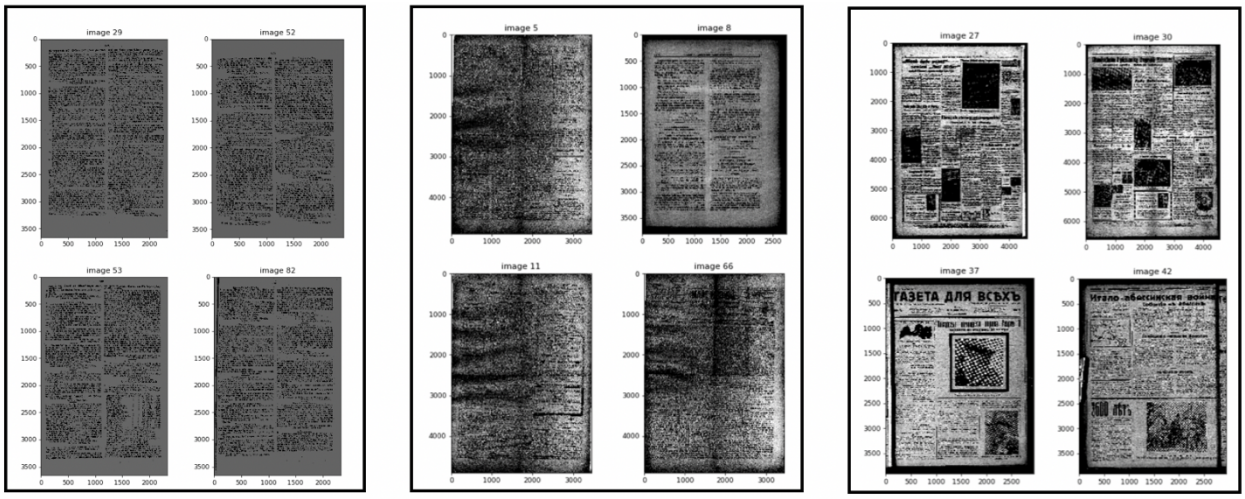
In the first part of this work, a visual inspection of sampled images from the same clusters does suggest shared visual features; for example, all four images in each box in Figure 15 show similar degrees of brightness and contrast. This result implies that there is a certain amount of intraclass correlation; images in the same cluster somewhat resemble each other.



(A) (B) (C)

FIGURE 15. IMAGES FROM THREE DIFFERENT CLUSTERS. IMAGES IN THE SAME CLUSTER SHARE SIMILAR CHARACTERISTICS, WHEREAS OTHER CLUSTERS SHOW DIFFERENT CHARACTERISTICS. FOR EXAMPLE, IMAGES IN (A) SHOW HIGH CONTRAST AND SIMPLE LAYOUT STRUCTURE. THE IMAGES IN (B) SHOW A RELATIVELY GRAY APPEARANCE WITHOUT FIGURE COMPONENTS. THE IMAGES IN (C) SHOW A RELATIVELY DARKER APPEARANCE WITH FIGURE COMPONENTS.

Following the first part of this exploration, we questioned whether the clustering results were simply based on the intensity value of the images. Thus, in the second part of this exploration, we clustered deep visual representations extracted from images that have been normalized to have zero mean and a unit standard deviation of intensity value. See Figure 16. From this second phase of the exploration, we observe two things. First, the clustering result using the deep visual representation excluding intensity shows a similar pattern to that of using the deep visual representation including intensity. This outcome suggests that the performance of our clustering approach is not based primarily on intensity features. Second, based on the observation that some datapoints sharing similar layout structures are slightly separated from each other compared to the first experiment clustering result, intensity does have an effect on the clustering process, even if it is not a primary basis of clustering.



(A) (B) (C)

FIGURE 16. IMAGES FROM THREE DIFFERENT CLUSTERS FOLLOWING INTENSITY VALUE NORMALIZATION. THE RESULTS SHOW SIMILAR CLUSTERING PATTERNS AS IN THE PREVIOUS FIGURE.

This exploration suggests that a set of deep visual representations of document images can be mapped into a low-dimensional space efficiently and effectively and neighboring datapoints show considerable visual similarity. This visual similarity is not based primarily on simple intensity features but rather on high-level visual features, such as layout density. We see potential in document clustering for enriched metadata as well as in using visually similar images for launching further study of materials within the Library of Congress and for allowing researchers/users to see visually similar images to those they are currently exploring.

For future development, we recommend looking at unsupervised approaches in order to build a more generic clustering solution not limited to a particular document domain or corpus. We also recommend exploring more sophisticated approaches to reducing dimensionality than what we have adopted here, in order to retain spatial information. And, we imagine combining this clustering approach with the results of document image quality assessment and the notion of a document complexity score, in order to see if there is a correlation between image quality assessment, complexity, and clustering.

7 DISCUSSION

The explorations framed above only scratch the surface of the types of investigations to be pursued with machine learning and the information that can be gleaned from and about digitized materials, the collections in which they sit, and about organizational and institutional practices and beliefs. We knew from the outset of this demonstration project that scoping our work would be crucial; we were already aware of the magnitude of possibilities and potential investigations, whether for a short-term project such as this one or more sustained research and development. Nonetheless, through the above explorations, we developed a heightened awareness of the number of possibilities and challenges, both those social and technical, as well as of their scale. In this section, we move from the more specific questions and narrower areas of focus pursued in the explorations to a set of themes, ideas, and questions that served as a backdrop to or emerged over the course of the larger project.

7.1 SOCIAL

Processing image and textual data with existing machine learning platforms and programs is increasingly accessible. That is not to say that doing so is exactly plug-and-play, but the technical, conceptual, and domain knowledge needed to simply pass data in to a machine learning pipeline and obtain some results appears lower each year. This perceived simplicity, however, hides significant complexity, nuance, assumptions and decision-making, and labor. Furthermore, this perceived simplicity has the potential to mask the implications of machine learning-generated knowledge, implications which range from the humorous and mundane to the profound and life-changing.

Domains considering implementing machine learning must engage deeply and critically with the technology, what it does, and what it means. For cultural heritage digital libraries, now is a critical moment to grapple with epistemologies of machine learning and the knowledge it structures, shapes, and appears to codify. Some elements of these epistemological conversations may transcend domains and applications, but these conversations also must be rooted in the specificities of the cultural heritage sector. In particular, libraries must grapple with their historical foundations and practices and with the potential consequences of these practices for machine learning. Previous and ongoing collecting and description practices, for example, were and are colonialist, racist, hetero- and gender-normative, and supremacist in other structural and systemic ways. These understandings are the foundation on which training and validation data will be created and assembled; they will become reinscribed as statements of truth, even as we elsewhere champion the potential of computational approaches to uncover hidden histories, identities, and perspectives in collections. To engage machine learning in cultural heritage must mean confronting these histories, committing to the hard work of acknowledgment and rectification, and not simply reproducing them and giving them a whole new scale of power. There should not be a future for machine learning in digital libraries that is not first and foremost committed to, in the words of Thomas Padilla, "responsible operations" and to all of the ongoing, cross-cutting work that responsible operations entail.[17]

Early in this demonstration project, Meghan Ferriter framed a range of different types of machine learning explorations and their outcomes. These included machine learning in the Library of Congress for description, discovery, and delight.[18] Ferriter's framing highlights another important feature in considering machine learning for digital libraries. Each of these endeavors--machine learning for description, discovery, and delight—has the potential to help people see materials from new angles, to peruse them in alternative ways, and to begin to frame additional questions and ways of thinking. At the same time, each of these purposes foregrounds different values and carries with it a different set of requirements and responsibilities. Naming and framing such purposes can help us think about the requirements and responsibilities of projects with these different ends. Building on Ferriter's "three Ds," we add as well "deployment" and "debate/dialogue." These categories need not be mutually exclusive, nor do we imagine a prescription for how machine learning in any of these realms should proceed. Instead, as a community of practice and as communities of researchers, what do we expect from projects and applications that proceed with these—and other—purposes in mind? Perhaps most critically, for any project that is about large-scale deployment, or a deployment of machine learning that may have significant implications for reasons beyond scale, what expectations do we hold as to what such projects must do, consider, make transparent? What contexts must we be able to see and understand?

7.2 TECHNICAL

Across the above explorations, several high-level themes and questions routinely surfaced related to what we might consider technical aspects of supporting and pursuing machine learning in the Library of Congress.

[17] Padilla, *Responsible Operations*.
[18] Ferriter, et al., Kick-off meeting.

The first is that on a basic level, computational access to the Library of Congress's digital objects is relatively straightforward. We were able to retrieve significant data—image, textual—via the Library's application programming interface and other bulk download options. This collections as data approach is an important layer for machine learning.

Even with this relatively straightforward baseline access to the digital materials, however, we depended on our inside access to people at the Library, made possible through this demonstration project, in order to make sense of some of the data. For example, what does the difficulty score encoded in the Beyond Words JSON data mean? How was it determined and by whom? How do the coordinates and size information in Chronicling America OCR.xml files correspond with size and coordinate information in Beyond Words? Are there existing values for digitization source type for digital collections and items? *Such questions and challenges may suggest the need for additional levels of documentation and/or to new types of reference support needed in the Library of Congress as it facilitates emergent areas of research with its digital collections.* We anticipate that the Library's Mellon-funded project, Computing Cultural Heritage in the Cloud, will further advance thinking and conversations on these topics.

Basic, computational access to digital collections ticks one box in the roadmap toward machine learning. Machine learning approaches also require accurate ground truth data from which to learn and validate. In the case of the explorations framed above, even when it seemed we could utilize existing Library of Congress data—generated by in-house experts or members of the public through crowdsourcing—as ground truth, *ground truth data proved challenging*. Ultimately, we had to create ground truth sets ourselves or turn to externally available datasets that provided the type/nature of ground truth information needed. Sometimes, we had to create these ground truth sets because the data did not otherwise exist as verifiable data, as in the case of the handwritten-typed-mixed project, for example. In other cases, the nature of the ground truth did not fit with our proposed approach, as in the difference between the rectangular bounding boxes of the Beyond Words project and the shape-fitting segmentation of our efforts in document segmentation and graphical content extraction. This reality about ground truth data was not wholly unexpected and is not a criticism of the Library's efforts or of individuals' labor and effort over time. What it may suggest, however, is that the *bibliographic information and collections-centered metadata previously pursued in libraries is a limited vision of what will be needed for machine learning applications and new areas of research*.

The lack of robust, varied, sizable, well-documented ground truth is a significant technical challenge to the development of machine learning approaches for cultural heritage. Cultural heritage digital libraries need ground truth data particular to their types of materials and also relevant to the type and variety of questions information professionals, researchers of various domains, and other users wish to pursue about these materials. In broad strokes, machine learning models developed and trained on other types of ground truth sets skew toward the contemporary and born-digital and are transferable to digitized historical materials only to a point. Maringanti, Samarakoon, and Zhu report, for example, that the current learning models for photograph description have been developed on photographs of the modern world and do not fit well with historical photographs in a research library's collections.[19] Likewise, historical materials may introduce additional challenges of noise and quality, whether due to material conditions, legacies of care, intervening technological processes, and more. Furthermore, datasets for competitions that focus on historical documents are relatively small, they are not comprehensive of the range of materials in collections as large and diverse as those in cultural heritage institutions, the statements about ground truth represented in them are typically narrow in application, and such ground truth sets are often siloed.

[19] Maringanti, Samarakoon, and Zhu, "Machine Learning Meets Library Archives."

The challenges around ground truth connect with other questions that surfaced across many of our explorations. These questions included, how might data created by users via the Library of Congress's crowdsourcing projects be used as ground truth data? What size of ground truth and training sets are necessary for different purposes? Are ground truth data created for one purpose transferrable for other purposes? What happens when we attempt to extrapolate from ground truth created for one purpose to another? Or when there isn't a direct match between ground truth data and output data?

Likewise, as a backdrop to a number of the explorations, we wondered about *the interplay of human expertise and processes and machine knowledge and processes.* What human-computer processes might be viably and validly adopted and operationalized as, say, part of a daily routine? What human-computer approaches are viable and valid in terms of effectiveness and efficiency in order to address issues of scalability? What value might there be in cross-learning, loop-learning, and cross-processing, where machines learn from humans, humans respond to and adapt understanding based on machine learning, and this looped learning informs processes and decision-making? Rather than seeing machine learning as an end, how can the Library of Congress embed and value critique across such a system, so that both human and machine assumptions are routinely tested? What are the foundational data and metadata needed and required to facilitate cross-learning and cross-processing? What is the place for data-science paradigms, where problems or issues are derived bottom-up—are surfaced through the collections and feature analysis—rather than top-down? We would be premature and ill-equipped to answer such questions based on only the explorations above, but we highlight them here as recurrent questions and provocations over the course of the explorations.

7.3 Social-Technical

In separating the technical from the social above, our intention is not to suggest a binary, a neat division, or a siloing of responsibility.[20] There is a critical interplay between the social and the technical in machine learning, not least because machine learning has significant social consequences. At best, machine learning will be incomplete without the social and cannot proceed without the technical, and technology cannot be divorced from the societies and individuals that develop it or the social realities it constructs. Our distinction above is largely one of convenience, and the lines between the social and technical blur quickly in our recommendations.

8 Recommendations

As the largest library in the world and with the ambitious, forward-looking digital strategy announced in 2019, the Library of Congress is uniquely situated to play a leadership role in advancing the theory and practice of machine learning in the cultural heritage sector. With this leadership role in mind, we propose two top-level recommendations for the Library of Congress's efforts around machine learning and as it moves forward in its work to "throw open the treasure chest," "connect," and "invest in our future."[21] The first is that the Library should focus the weight of its machine learning efforts and energies on social and technical infrastructures for the development of machine learning in cultural heritage organizations, research libraries, and digital libraries. Second, we recommend that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections.

[20] Padilla refers to the "technical, organizational, and social challenges" as multiple, integral facets of a machine learning agenda for libraries. See *Responsible Operations*, p. 6.
[21] Library of Congress, "Digital Strategy for the Library of Congress."

What we do *not* recommend at this particular moment in time is the broad application of machine learning to the Library's digital collections with the purpose of broadly making claims about the materials or restructuring access to them. On a very practical level, such broad application would be premature due to the challenges with ground truth data and validation and considering the many critical conversations yet to take place around whose, and what, human knowledge becomes the basis of machine learning. We advise against a "more product, less process" approach to machine learning applications. The ways in which machine learning-generated knowledge stands to influence decision-making and codify particular understanding are too profound and too powerful to adopt such an approach, or make such a commitment, at this nascent stage.

Below, we provide more detail about the two top-level recommendations and offer several short- and medium-term recommendations in support of these top-level recommendations. Each of these top-level recommendations directly map to the three goals of the "Digital Strategy for the Library of Congress." In addition, while we hope the role and importance of people will be clear in everything that follows, we want to say directly here: people are central to all of the recommendations that follow. None of the recommendations imagine a library without information professionals and experts. Any future for machine learning in libraries will require an investment in people with many types of expertise, and a best-case future for machine learning in cultural heritage organizations is that the people who work in them are able to bring even more of their experience and expertise to bear.

8.1 FOCUS THE WEIGHT OF THE LIBRARY'S MACHINE LEARNING EFFORTS AND ENERGIES ON SOCIAL AND TECHNICAL INFRASTRUCTURES FOR THE DEVELOPMENT OF MACHINE LEARNING IN CULTURAL HERITAGE ORGANIZATIONS, RESEARCH LIBRARIES, AND DIGITAL LIBRARIES.

We recommend that the Library dedicate itself to a range of infrastructure projects that will create a strong foundation for machine learning in the profession and field, particularly as applied to historical cultural heritage materials. The paramount machine learning need within the cultural heritage sector at this time is the development of infrastructure.[22] These infrastructures include educative infrastructures, through which cultural heritage professionals develop further literacy in computational thinking and methods, particularly through the lens of critical information studies. The needed infrastructures also include platforms for conversations—and the pursuant conversations themselves—about the language of description and the corresponding social and cultural values signaled in that language, as well as about who we engage in these processes and applications. Likewise, the needed infrastructures include pathways for gathering and delivering machine learning models and verifiable learning data that extend beyond individual projects, as well as for bringing together cross-domain researchers toward the purpose of machine learning for cultural heritage.

This top-level recommendation, that the Library focus the weight of its energies on social and technical infrastructures, connects with many of the goals in the Library's digital strategy. This recommendation aligns particularly well with the Library's interests in *maximizing use of content*; *supporting emerging styles of research*; *welcoming other voices*; *driving momentum in our communities*; *cultivating an innovation culture*; *ensuring enduring access to content*; and *building toward the horizon*. (See Table 9, which maps our recommendations to the Library of Congress digital strategy.)

Under this broad umbrella, we propose several more specific recommendations, through which we believe the Library of Congress could have the most immediate and most significant impact, drawing on both its status and position as well as existing areas of expertise. We have not attempted to be exhaustive in these recommendations;

[22] Padilla has framed this need as a research agenda for libraries, including the areas of machine learning, data science, and artificial intelligence. See Padilla for an even more extensive list of needs.

instead we emphasize activities that we consider critical or high priority given community needs, the Library's existing expertise/leadership, and opportunity. To that end, we recommend that the Library of Congress should:

- Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage.
- Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to the larger cultural heritage and other research communities.
- Focus efforts on developing ground truth sets and benchmarking data and making these easily available.

In the following subsections, we describe these recommendations in more detail. We also map each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*. (See Table 10, which maps each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*.)

8.1.1 Develop a statement of values or principles that will guide how the Library of Congress pursues the use, application, and development of machine learning for cultural heritage.

The Library of Congress should articulate a statement of values or principles with regard to the adoption, use, and development of machine learning. Such a statement can address machine learning and cultural heritage broadly—what is the vision of machine learning for cultural heritage that Library of Congress aspires to—and also frame the values and principles under which the Library of Congress will pursue the development and application of machine learning. If units within the Library seek to apply machine learning to collections, under what principles and values should that work proceed? What are the expectations around transparency and explainability, both for internal and external audiences, for example? Or around confronting problematic historical knowledge and knowledge structures in training data? The crafting of such a statement of values or principles is an opportunity for developing increased literacy and fluency around machine learning, if the Library engages its staff broadly in the developing and education around such a statement.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
    - Managing Bias
    - Transparency, Explainability, Accountability
    - Distributed Data Science Fluency

8.1.2 Create and scope a machine learning roadmap for the Library that looks both internally to the Library of Congress and its needs and goals and externally to cultural heritage and other communities of research and practice.

For this demonstration project, we scoped our explorations above as a response to the seemingly endless array of opportunities for applying machine learning to the Library of Congress's digital collections. As the Library of Congress continues to explore the intersection of machine learning and digital cultural heritage, the Library likewise will need to focus and scope its efforts. Such scoping and an overall roadmap are necessary for the Library to influence and have a strong impact on the development of machine learning in cultural heritage organizations.

The roadmap should be informed by the statement of values recommended in 4.1.1. In addition, other recommendations in this report may be points on that roadmap. The investigation areas in *Responsible Operations* may provide a useful framework for such a roadmap. What are the Library's goals and objectives in each of the

investigation areas? Will it pursue all of the areas or prioritize particular areas? With regard to the Library's goals and objectives, are there investigations areas that the Library would add? Do recommendations from that report offer ways of thinking about and structuring a longer-term roadmap?

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
    - Transparency, Explainability, Accountability
    - Distributed Data Science Fluency
- Workforce Development
    - Investigating Core Competencies
    - Committing to Internal Talent
- Any/all of the investigation and challenge areas that the Library of Congress would choose to prioritize in its own roadmap and plan

8.1.3 Focus efforts on developing ground truth sets and benchmarking data and making these easily available.

One key way for the Library of Congress to advance machine learning for cultural heritage is creating and distributing ground truth sets drawn from its diverse digital collections and making available benchmarking data for computational approaches on those sets. Ground truth data and benchmarks will allow researchers—including cultural heritage professionals, computer scientists, and developers—to focus their energies and research, development, and analysis, rather than on creating one-off, niche datasets. The availability of ground truth and benchmarks also create the possibility of more rapid development around particular problem domains.

Creating and distributing ground truth sets will foreground the significance of metadata, including technical, structural, and descriptive. For descriptive metadata, we recommend distinguishing between at least two types of descriptive metadata, one that is descriptive of the content of the historical materials, including metadata about what is depicted and represented as well as how, and another that is descriptive of the properties of the image, including features such as digitization source, contrast, skew, noise, range effect, complexity (or a difficulty measure of some sort). Underscoring this idea is that the ground truth sets will have interest to researchers of many disciplines and research interests (those interested in the materials themselves as cultural objects and those interested in them for their value to computer science development, for example).

Within this recommendation, we offer two sub-recommendations:

8.1.3.1 Development of DocuNet

We recommend the Library of Congress develop, or partner in developing, DocuNet, an image database of historical documents with accompanying taxonomic and typological metadata. DocuNet would be valuable to researchers in library and related sectors and also to information science and computer science researchers. We see it as one effective way to encourage additional machine learning researchers and those interested in computer vision, among other domains, to delve into historical document analysis. Features or characteristics important to a DocuNet are ground-truth (e.g., document types, coordinates of article regions, etc.); openness; diversity and balance (e.g., different document types should be comprehensively covered and equally distributed); and clear objectives (e.g., segmentation, classification, clustering, etc.).

8.1.3.2 Pursuit of Low-cost Ground-truthing

We also recommend that the Library explore options for, and contribute to efforts to advance, low-cost ground-truthing. Having subject matter experts hand-label data is expensive and is a barrier to machine learning, whether in the Library of Congress or on external research teams. Importantly, by low-cost ground-truthing, we do not mean exploitative labor such as through Mechanical Turk. Instead, the Library could pursue heuristics-based models. In this form of ground-truthing, computers learn the heuristics that are created by humans, and then computers label data using the heuristic rules, constraints, distributions, and/or variances of the dataset. Such an approach may be less accurate than item-by-item expert-labeled ground truth, but it may still be able to produce effective machine learning systems. While the potential for low-cost ground-truthing of this sort remains to be seen, we believe it is a worthwhile area of inquiry as part of a larger commitment to support for machine learning infrastructure.

Recommendation 4.2.3 and its corresponding sub-recommendations map to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
    - Managing Bias
- Description and Discovery
    - Enhancing Description at Scale
- Shared Methods and Data
    - Shared Development and Distribution of Training Data

8.2 INVEST IN CONTINUED, ONGOING, INTENTIONAL EXPLORATIONS AND INVESTIGATIONS OF PARTICULAR MACHINE LEARNING APPLICATIONS TO ITS COLLECTIONS.

Much as it did through this demonstration project and other activities (its Innovator-in-Residence program, for example), the Library should continue to invest in explorations and investigations of particular applications of machine learning on its collections, with an eye toward both internal operations and impacts on external users. Continued explorations, tests, and experiments will prove crucial to the ongoing inquiry needed to more fully evaluate the potential of machine learning for digital libraries. We recommend that such explorations are framed and understood as *intellectual endeavors* rather than being large output-driven and are collaborations among computer scientists, developers, and information professionals, drawing in other participants and stakeholders as appropriate to the project. We also encourage the Library of Congress to be careful in the presentation of machine learning generated data, particularly when that data might be read or experienced by others as uncontested knowledge or fact about cultural heritage materials, and also with care and concern about what is absent as well as what is present.

This recommendation, that the Library invest in continued, ongoing, intentional explorations and investigations of particular machine learning applications to its collections, connects with many of the goals in the Library's digital strategy. These include, in particular, *supporting emerging styles of research*, *welcoming other voices*; *driving toward momentum in our communities*; *cultivating an innovation culture*; and *building toward the horizon*.

Again, we propose several more specific recommendations, through which we believe the Library of Congress could have the most immediate and most significant impact, drawing on both its status and position as well as existing areas of expertise. We have not attempted to be exhaustive in these recommendations; instead we emphasize activities that we consider critical or high priority given community needs, the Library's existing expertise/leadership, and opportunity. To that end, we recommend that the Library of Congress should:

- Join the Library of Congress's emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate.
- Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas.
- Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations.

In the following subsections, we describe these recommendations in more detail. We also map each recommendation to areas of investigation and challenge areas outlined in *Responsible Operations*.

8.2.1 Join the Library of Congress's Emergent efforts in machine learning with its existing expertise and leadership in crowdsourcing. Combine these areas as "informed crowdsourcing" as appropriate.

Through its By the People application and campaigns, as well as through earlier efforts, the Library of Congress has established a strong portfolio of crowdsourcing experience. Through the Library of Congress Labs, the Library also has strong leadership in crowdsourcing, including technical development for crowdsourcing, and in designing and developing challenges. We see significant potential in bringing together machine learning and crowdsourcing efforts, as an effort to combine an existing strength of the Library with an emergent area of interest and impact. Such a joining of efforts would allow the Library of Congress to make even greater use of crowdsourced information, toward challenges of scalability. Doing so also creates the opportunity for greater conceptual understanding and practical development. For example, joining these areas, even in a limited way, would allow the Library to research cross-learning and looped learning. Such a combined approach has the potential to improve machine learning models, particularly in applications that require a higher-level of understanding. In a hypothetical project, members of the crowd might receive labeled data from a model; users then revise the labels, and the model improves its predictions based on those revisions. With each successive iteration, the model improves further.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Description and Discovery
  - Enhancing Description at Scale
- Workforce Development
  - Committing to Internal Talent
- Shared Methods and Data
  - Shared Development and Distribution of Training Data

8.2.2 Sponsor challenges for teams to create additional metadata for digital collections in the Library of Congress through Machine Learning. As part of these challenges, require teams to engage across a range of social and technical questions and problem areas.

The Library has a history of creating and sponsoring challenges, such as the Congressional Data Challenge and challenges focused on Chronicling America data. We recommend that the Library build on this prior experience to organize and offer new sponsored challenge opportunities about machine-learning generated metadata. Such explorations have the potential to move forward a range of critical conversations and needs. The purpose of this recommendation is multipart: (1) To see what types of metadata researchers/teams might produce. What metadata is of interest to them? (2) To encourage the creation of particular types of metadata, including through an expanded sense of what descriptive metadata might include and what is of descriptive value (e.g., metadata that are

representational of the content and metadata that are descriptive of features such as noisiness, quality, and so on); (3) To anchor critical engagement with core problems, such as of bias in the data and in what may be produced, as inseparable from technical development; and (4) To emphasize, underscore, and champion that cross-disciplinary, community-centered and community-engaged development is required for responsible machine learning.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Committing to Responsible Operations
  - Managing Bias
  - Transparency, Explainability, and Accountability
- Description and Discovery
  - Enhancing Description at Scale
- Shared Methods and Data
  - Shared Development and Distribution of Methods
- Sustaining Interprofessional and Interdisciplinary Collaboration

8.2.3 Continue to create and support opportunities for researchers to partner in substantive ways with the Library of Congress on machine learning explorations.

Even if the Library were able to dedicate many staff members to a full-time focus on machine learning, the challenges of machine learning for cultural heritage are large and significant enough that the Library will need to continue its collaborations with external researchers. Such opportunities need not be sponsored by the Library itself, though they could be. However they are facilitated, we recommend that the Library see formal collaborations as central to taking this machine learning work forward. As researchers who have worked with Library of Congress data for many years—and over which time we have had many positive and helpful interactions with Library staff, who probably went well beyond the call of duty in their help to us—we benefitted in significant ways from the additional levels of access to Library staff this this particular demonstration project and the formal collaboration afforded. Understandably, the Library cannot support every machine learning endeavor at this level or engage with every research team in this way, and there will be challenges of scale. Nonetheless, we recommend that some measure and shape of formal collaboration opportunities be part of the Library's support for both machine learning explorations and larger social and technical infrastructures.

This recommendation maps to the following investigation areas and challenges in *Responsible Operations*:

- Shared Methods and Data
  - Shared Development and Distribution of Methods
- Sustaining Interprofessional and Interdisciplinary Collaboration

TABLE 9. INFRASTRUCTURE AND APPLICATION RECOMMENDATIONS MAPPED TO ELEMENTS OF THE LIBRARY OF CONGRESS'S DIGITAL STRATEGY.

Digital Strategies	Recommendations on Infrastructure	Recommendations on ML Applications
maximizing use of content	✓	
supporting emerging styles of research	✓	✓
welcoming other voices	✓	✓
driving momentum in our communities	✓	✓
cultivating an innovation culture	✓	✓
ensuring enduring access to content	✓	
building toward the horizon	✓	✓

TABLE 10. RECOMMENDATIONS MAPPED TO AREAS OF INVESTIGATION AND CHALLENGE AREAS OUTLINED IN PADILLA'S RESPONSIBLE OPERATIONS.

Strategies	Sub-Strategies	Statement of Vision	Roadmap of ML	Ground-Truthing & Benchmarking	ML + Crowd-sourcing Efforts	Sponsoring Challenges	Research Partnerships
Committing to Responsible Operations	Managing Bias	✓	✓	✓		✓	
	Transparency, Explainability, Accountability	✓	✓			✓	
	Distributed Data Science Fluency	✓	✓				
Workforce Development	Investigating Core Competencies		✓		✓		
	Committing to Internal Talent		✓				
Description & Discovery	Enhancing Description at Scale			✓	✓	✓	
Shared Methods and Data	Shared Development and Distribution of Training Data			✓	✓		
	Shared Development and Distribution of Methods					✓	✓
Sustaining Interprofessional & Interdisciplinary Collaboration						✓	✓

9 CONCLUSION

There is rich potential for machine learning to augment the description and accessibility of materials in the Library of Congress, to inform understanding of collections and choices about how materials are processed and by whom, and to address issues of scale. The Library of Congress is in a remarkable position to advance machine learning for cultural heritage organizations, through its size, the diversity of its collections, and its commitment to digital strategy. This demonstration project—via its explorations, discussion, and recommendations—has shown the potential of machine learning toward a variety of goals and use cases, and it has argued that the technology itself will not be the hardest part of this work. The hardest part will be the myriad challenges to undertaking this work in ways that are socially and culturally responsible, while also upholding responsibility to make the Library's materials available in timely and accessible ways.

BIBLIOGRAPHY

Afzal, Muhammad Zeshan, Andreas Kölsch, Sheraz Ahmed, and Marcus Liwicki. "Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification." In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:883–88, 2017. https://doi.org/10.1109/ICDAR.2017.149.

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. "dhSegment: A Generic Deep-Learning Approach for Document Segmentation." In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12, 2018. https://doi.org/10.1109/ICFHR-2018.2018.00011.

Ferriter, Meghan, et al. Kick-off meeting for "Digital Libraries, Intelligent Data Analytics, and Augmented Description." Washington, DC, July 16, 2019.

Harley, Adam W., Alex Ufkes, and Konstantinos G. Derpanis. "Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval." In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 991–95, 2015. https://doi.org/10.1109/ICDAR.2015.7333910.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition," 770–78, 2016. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kong, Tao, Anbang Yao, Yurong Chen, and Fuchun Sun. "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection," 845–53, 2016. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Kong_HyperNet_Towards_Accurate_CVPR_2016_paper.html.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Curran Associates, Inc., 2012. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Library of Congress. "Digital Strategy for the Library of Congress." 2019. https://www.loc.gov/digital-strategy/

Lorang, Elizabeth, Leen-Kiat Soh, Yi Liu, Chulwoo Pack, and Delaram Rahimi, "Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures," National Digital Newspaper Program Annual Meeting, Washington, D.C., September 26, 2018. https://digitalcommons.unl.edu/library_talks/143/

Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9, no. Nov (2008): 2579–2605.

Maringanti, Harish, Dhanushka Samarakoon, and Bohan Zhu, "Machine Learning Meets Library Archives: Image Analysis to Generate Descriptive Metadata." 2019. https://www.lyrasis.org/Leadership/Documents/Catalyst%20Fund/UU-version2-MachineLearning-CatalystFund-WhitePaper.pdf

Odena, Augustus, Vincent Dumoulin, and Chris Olah. “Deconvolution and Checkerboard Artifacts.” *Distill* 1, no. 10 (October 17, 2016): e3. https://doi.org/10.23915/distill.00003.

Padilla, Thomas. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. 2019. https://doi.org/10.25333/xk7z-9g97.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, edited by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, 234–41. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. “ImageNet Large Scale Visual Recognition Challenge.” International Journal of Computer Vision, 2015.

Seguin, Benoit and Sofia Ares Oliveira. dhSegment [computer program]. https://github.com/dhlab-epfl/dhSegment.

Simonyan, Karen, and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *ArXiv:1409.1556 [Cs]*, April 10, 2015. http://arxiv.org/abs/1409.1556.

Xie, Saining, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. “Aggregated Residual Transformations for Deep Neural Networks,” 1492–1500, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Xie_Aggregated_Residual_Transformations_CVPR_2017_paper.html.

Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. “EAST: An Efficient and Accurate Scene Text Detector,” 5551–60, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_EAST_An_Efficient_CVPR_2017_paper.html