Digitizing and parsing semi-structured historical administrative documents from the G.I. Bill mortgage guarantee program

Sara Lafia (slafia@umich.edu), Research Fellow[1]
David A. Bleckley, Senior Data Project Manager[1]
J. Trent Alexander, Associate Director[1]

1. ICPSR, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Purpose

Many libraries and archives maintain collections of research documents, such as administrative records, with paper-based formats that limit their access to in-person use. Digitization transforms paper-based collections into more accessible and analyzable formats. As collections are digitized, there is an opportunity to incorporate deep learning techniques, such as Document Image Analysis (DIA), into workflows to increase the usability of information extracted from archival documents. This paper describes our approach using digital scanning, optical character recognition (OCR), and deep learning to create a digital archive of administrative records related to the mortgage guarantee program of the Servicemen's Readjustment Act of 1944, also known as the G.I. Bill.

Design/methodology/approach

We used a collection of 25,744 semi-structured paper-based records from the administration of G.I. Bill Mortgages from 1946 to 1954 to develop a digitization and processing workflow. These records include the name and city of the mortgagor, the amount of the mortgage, the location of the Reconstruction Finance Corporation agent, one or more identification numbers, and the name and location of the bank handling the loan. We extracted structured information from these scanned historical records in order to create a tabular data file and link them to other authoritative individual-level data sources.

Findings

We compared the flexible character accuracy of five OCR methods. We then compared the character error rate of three text extraction approaches (regular expressions, document image analysis, and named entity recognition). We were able to obtain the highest quality structured text output using DIA with the Layout Parser toolkit by post-processing with regular expressions. Through this project, we demonstrate how DIA can improve the digitization of administrative records to automatically produce a structured data resource for researchers and the public.

Originality/value

Our workflow is readily transferable to other archival digitization projects. Through the use of digital scanning, OCR, and DIA processes, we created the first digital microdata file of administrative records related to the G.I. Bill mortgage guarantee program available to researchers and the general public. These records offer research insights into the lives of veterans who benefited from loans, the impacts on the communities built by the loans, and the institutions that implemented them.

Keywords**:** archives, digitization, document image analysis, historical records, OCR, workflows
Paper type*:* research paper

Introduction

Digitization has been described as an activity in which information about objects and their context can be converged into a single system (Navarrete & Owen, 2011). Since 2004, digitization has been recognized as a preservation reformatting method by the Association of Research Libraries (ARL) for ensuring continued access to paper-based materials (Arthur et al., 2004). Beyond simply preserving materials, the decision to digitize a collection introduces new possibilities for enhancing access and description of collections. Well-known digitization projects, like Google's campaign to digitize books, have enabled the large-scale analysis of documents by providing new interactions, like zooming in on scanned images, and ways to access primary source materials online (Leetaru, 2008).

The digitization of paper-based materials is now standard practice in libraries, archives, and museums (Lischer-Katz, 2022). Technical guidelines for digitizing cultural heritage materials describe workflows in which materials are cataloged, scanned, reviewed for quality, archived, and published (Federal Agencies Digital Guidelines Initiative, 2022; Puglia et al., 2005). Efforts to embed technologies, like open-source optical character recognition (OCR), into digital historical research workflows (Blanke et al., 2012) have made digitization more relevant, accessible, and customizable for adoption by specific research communities. Emerging technologies like deep learning and image segmentation are poised to augment digitization workflows by capturing the structure and content of textual documents (Shen et al., 2021).

Much of the historical documentation OCR literature focuses on the digitization of prose documents or the conversion of hard copy tabular records into digital tabular data (Nagy, 1992; Stančić & Trbušić, 2020). However, bridging these use cases is the digitization of semi-structured historical documents, which hold data that could be converted into a tabular format but are not currently formatted into forms or tables. For example, individual records can be digitized and tabulated for large-scale analysis (Brahney, 2015). Document structures are also critical for maintaining the full meaning of documents, like newspapers, and provide valuable context for text mining and historical analysis (Lee et al., 2020). A movement to treat "collections as data" argues that as text is digitized into machine-actionable corpora, including document structure in the text digitization process enables computational research methods such as text mining, data visualization, mapping, and network analysis (T. Padilla et al., 2019).

This paper investigates the feasibility of augmenting a conventional workflow to digitize and parse semi-structured archival records using open-source document image analysis (DIA) and named entity recognition (NER) approaches. We applied these methods to digitize a collection of 25,744 paper-based records from the administration of mortgage guarantees by the Servicemen's Readjustment Act of 1944, commonly known as the G.I. Bill. The output of our process represents the first tabular administrative records dataset available for the study of the implementations and outcomes of the G.I. Bill mortgage guarantee program. We evaluated the performance of multiple OCR methods as well as each text extraction approach by comparing its output against ground truth data. We found that DIA, post-processed using regular expressions, produced the highest quality dataset of structured text. This paper contributes: 1) a digitization workflow for recovering structured text from administrative records; and 2) a novel data collection available to study the G.I. Bill mortgage guarantee program and its beneficiaries.

Background

Text extraction methods

The process of scanning paper-based documents produces representative digital images, which can be indexed for search and retrieval and distributed online. Scanning also enables the conversion of raster images into text through a process known as optical character recognition–or OCR (Stevens, 1961). OCR was originally developed on typewritten cards to support data entry from paper-based records (Leimer, 1962). Scanned digital images need to support OCR conversion and ensure a high quality of output text (Booth & Gelb, 2006).

Contemporary OCR approaches are more flexible than their predecessors in that they take advantage of document structure to process text blocks (e.g., captions, words in tables) instead of recognizing single characters at a time (Nagy, 1992). Leading OCR engines, like Tesseract, are capable of capturing scanned text with relatively high accuracy, provided that the documents have been correctly prepared and pre-processed (Smith, 2007). OCR engines perform text segmentation and character prediction through classification (Neudecker et al., 2021). Measures such as character accuracy and character error rate (Neudecker et al., 2021; Rice, 1996) are useful for determining the performance of OCR apart from other steps in digitization workflows. Despite the fact that OCR engines can extract scanned text with high accuracy, the output of OCR is unstructured. In other words, because OCR does not capture the layout context of documents (e.g., columns or fields), separate steps, such as layout analysis, are often needed to detect and evaluate the sources of errors in text extraction (Packer, 2011).

Document image and layout analysis

OCR is just one component of document processing workflows, which also includes page layout analysis and other document image analysis (DIA) techniques (Kasturi et al., 2002). Modern DIA methods take advantage of deep learning to classify images and detect document layouts. In recent years, deep learning methods using convolutional neural networks have advanced the state of the art for complex text digitization tasks, such as medieval handwriting classification and layout analysis (Pondenkandath et al., 2017). Strategies and datasets originally developed for computer vision research have been adapted for use in other domains through transfer learning. For instance, the PubLayNet dataset was originally trained to detect the layout of scientific articles and has provided a foundation for developing custom applications for layout detection and analysis of other sources of text (Zhong et al., 2019).

Detecting and incorporating the layout of documents into information extraction tasks makes it possible to retain the layout context of original documents. However, the procedures for adapting and tuning existing document image datasets can be complicated to reproduce due to the use of proprietary services or the need to manage numerous software dependencies. For example, a pipeline was recently developed for digitizing scanned card index records (Amujala et al., 2023). However, its reliance on proprietary OCR and natural language processing (NLP) services available through Amazon Web Services makes it difficult for other researchers to inspect or adapt the underlying models. By contrast, the LayoutParser toolkit supports document image processing and structured text extraction, enabling researchers to adapt existing image layout detection pipelines for custom text extraction tasks (Shen et al., 2021). Techniques like Named Entity Recognition (NER) can further improve the quality and structure of extracted text during post-processing. For example, NER can be used to predict tags for extracted entities (Lu et al., 2013) or recognize semi-structured entities (Irmak & Kraft, 2010).

Materials and Methods

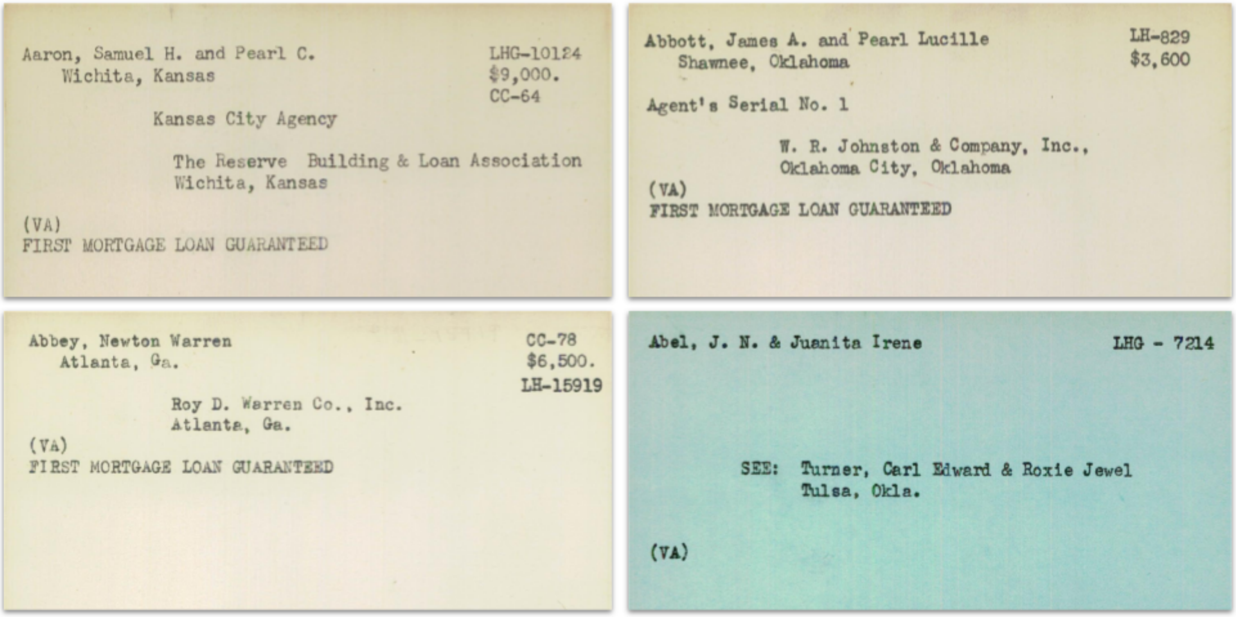
Mortgage record index cards

We focus on the digitization of paper-based records from the administration of the mortgage guarantee program of the Servicemen's Readjustment Act of 1944, known as the G.I. Bill, which guaranteed loans made to U.S. veterans of World War II (*Servicemen's Readjustment Act of 1944*, 1944). Between 1944 and 1952, the program guaranteed over two million mortgages (United States Department of Veterans Affairs, 2013). Prior studies examining the impact of the G.I. Bill have relied on indirect evidence; the literature provides no analysis of administrative records from the implementation of the program (Katznelson & Mettler, 2008).

The *Index to Loans on Veterans Administration Guaranteed Mortgages, 1946 – 1954 ("Index to Loans on Veterans Administration Guaranteed Mortgages, 1946 – 1954," n.d.)* is a collection of 23 linear feet of three-inch by five-inch index cards housed at the National Archives and Records Administration (NARA) in College Park, Maryland. Documents from this collection offer the first administrative data on the execution of the G.I. Bill available to researchers and the general public. Though they are not a comprehensive record of all mortgage guarantee beneficiaries, they constitute a large, novel dataset (n=25,744 scanned images) that is well-suited for analyzing the long-term impacts of the G.I. Bill program.

Each index card in the collection contains information about the name and address of the mortgagor, the amount of the mortgage, the name of the RFC loan agency approving the mortgage with the issuing agent's serial number, and the name and location of the bank handling the mortgage loan (Zaid, 1973). While most of the index cards contain these common fields, they are hand-typed, and the text fields are not in identical positions on each card. **Figure 1** illustrates the variation in card layouts. Given that the cards were hand-typed, we expected that OCR methods would convert scanned text to electronically-encoded characters with a high degree of accuracy. The variety of card layouts, however, presented challenges for maintaining the layout context during parsing, using both machine learning and standard approaches.

To evaluate the effectiveness of our text extraction approaches, we developed three truth decks. The first truth deck (n=100) contained hand-keyed text files for evaluating the quality of the OCR output text. The second truth deck (n=500) contained cards with regions that we labeled and used as training data for learning card layouts. We labeled batches of 100 cards at a time and trained the model iteratively until we saw only marginal gains in average precision per label category. **Figure 2** shows how we used the open-source software, LabelStudio (Tkachenko et al., 2020) to label the text fields *Name*, *Location*, *Amount*, *ID*, *Status*, *Agency*, and *Other*. These labels provided mappings between text fields and the spatial regions of the cards in which they tended to occur. The third truth deck (n=100) contained hand-keyed data assigned to specific fields to assess the accuracy of parsing.



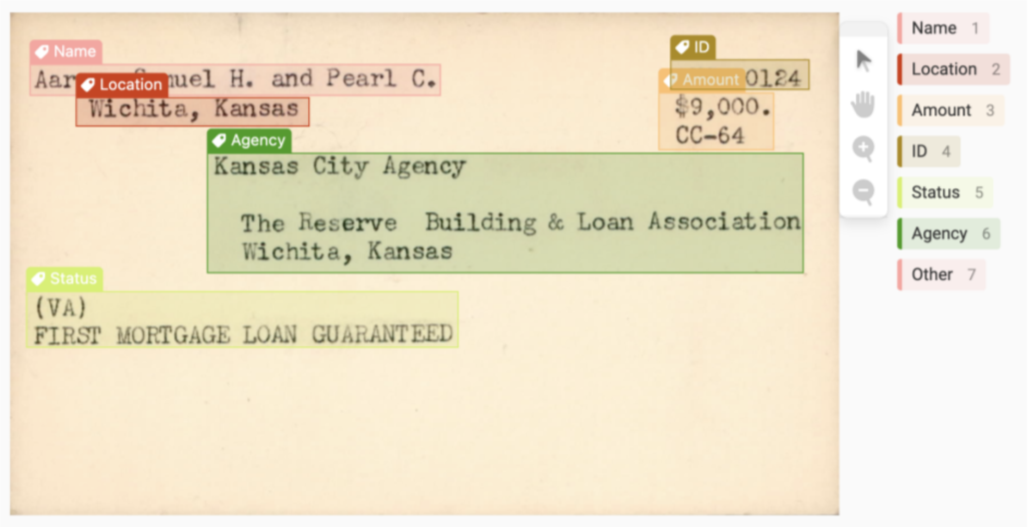
Aaron, Samuel H. and Pearl C.
Wichita, Kansas
LHG-10124
$9,000.
CC-64
Kansas City Agency
The Reserve Building & Loan Association
Wichita, Kansas
(VA)
FIRST MORTGAGE LOAN GUARANTEED

Abbott, James A. and Pearl Lucille
Shawnee, Oklahoma
LH-829
$3,600
Agent's Serial No. 1
W. R. Johnston & Company, Inc.,
Oklahoma City, Oklahoma
(VA)
FIRST MORTGAGE LOAN GUARANTEED

Abbey, Newton Warren
Atlanta, Ga.
CC-78
$6,500.
LH-15919
Roy D. Warren Co., Inc.
Atlanta, Ga.
(VA)
FIRST MORTGAGE LOAN GUARANTEED

Abel, J. N. & Juanita Irene
LHG - 7214
SEE: Turner, Carl Edward & Roxie Jewel
Tulsa, Okla.
(VA)

Figure 1. Example of four scanned mortgage record index cards from the Index to Loans on Veterans Administration Guaranteed Mortgages (1946 – 1954)



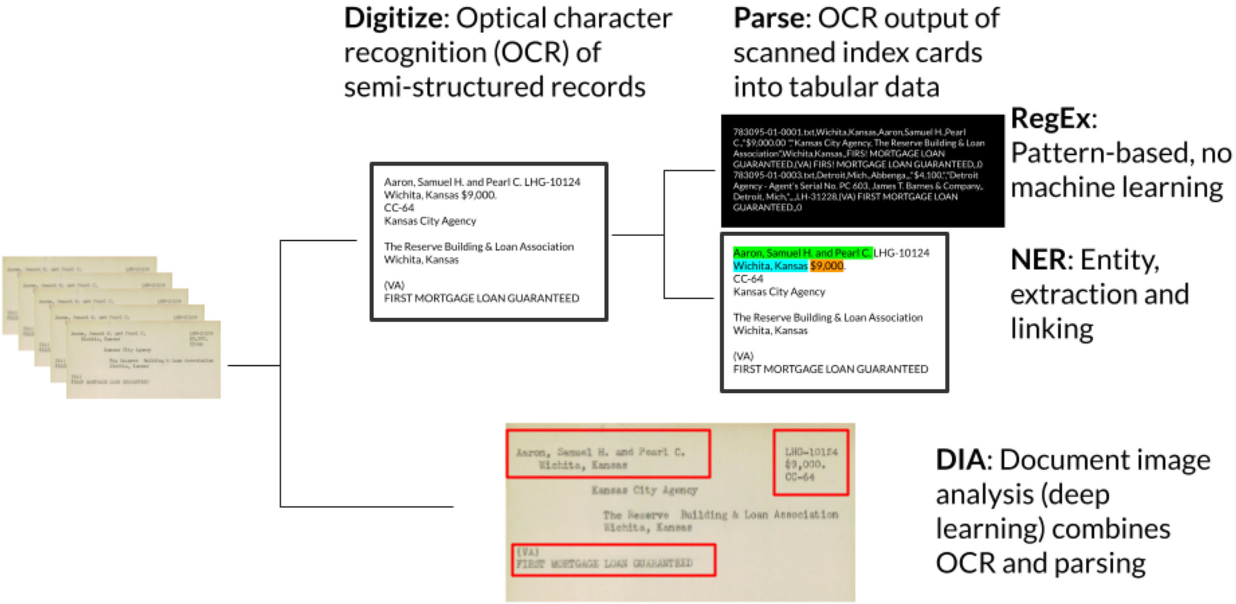
Aar muel H. and Pearl C.
Wichita, Kansas
0124
$9,000.
CC-64
Kansas City Agency
The Reserve Building & Loan Association
Wichita, Kansas
(VA)
FIRST MORTGAGE LOAN GUARANTEED

Figure 2. Labeling layout training data indicating text regions with LabelStudio software

Digitization and parsing workflow

Archival records are often stored as card indexes, which serve as finding aids or contain individual-level information. Historically, the digitization of index cards has been a manual process subject to human error (e.g., inconsistent data entry) and computational error (e.g., inaccurate character recognition) (Amujala et al., 2023). A key challenge of extracting structured textual information from semi-structured historical records is incorporating their layout information into digitization workflows (Shen et al., 2021).

To address this challenge, we employed and assessed multiple methods of digitizing and parsing index card images to develop a workflow that leverages the layout of scanned cards to extract and structure their text. **Figure 3** provides an overview of the digitization and parsing methods we used and evaluated in creating our workflow. The workflow transforms the paper-based records into a combined tabular data file, suitable for record linkage and historical analysis. The final model segments each scanned card into regions and predicts a labeled bounding box for each corresponding block of text.


Digitize: Optical character recognition (OCR) of semi-structured records
Parse: OCR output of scanned index cards into tabular data
RegEx: Pattern-based, no machine learning
NER: Entity, extraction and linking
DIA: Document image analysis (deep learning) combines OCR and parsing

Aaron, Samuel H. and Pearl C. LHG-10124
Wichita, Kansas $9,000.
CC-64
Kansas City Agency

The Reserve Building & Loan Association
Wichita, Kansas

(VA)
FIRST MORTGAGE LOAN GUARANTEED

```
783095-01-0001.txt,Wichita,Kansas,Aaron,Samuel H.,Pearl
C.,"$9,000.00 ","Kansas City Agency, The Reserve Building & Loan
Association",Wichita,Kansas,,FIRS! MORTGAGE LOAN
GUARANTEED,(VA) FIRS! MORTGAGE LOAN GUARANTEED,,0
783095-01-0003.txt,Detroit,Mich.,Abbenga,,"$4,100.","Detroit
Agency - Agent's Serial No. PC 603, James T. Barnes & Company,
Detroit, Mich.",,,LH-31228,(VA) FIRST MORTGAGE LOAN
GUARANTEED,,0
```

Aaron, Samuel H. and Pearl C. LHG-10124
Wichita, Kansas $9,000.
CC-64
Kansas City Agency

The Reserve Building & Loan Association
Wichita, Kansas

(VA)
FIRST MORTGAGE LOAN GUARANTEED

Figure 3. Workflow for digitizing and parsing the scanned historical documents

In the first step, *digitization*, the cards were scanned to high-resolution digital images using flatbed scanners by staff at NARA. We then used optical character recognition (OCR) to extract variably-structured text from the scanned images. To identify the best method for OCR, we tested five standard methods (see **Table 1**). These included both stand-alone software – (ABBYY FineReader PDF (ABBYY, 2019), Acrobat Pro (Adobe, 2022), and OmniPage Professional (Nuance Communications, Inc., 2011) – and OCR engines called within Python workflows – Tesseract used in LayoutParser (Shen et al., 2021) and Python-tesseract (Hoffstaetter, 2021; Tesseract, 2021). We evaluated the quality of the OCR using the hand-keyed truth deck that we created (see **Table 4** in the Results section). In addition to the digital images from NARA, the first step output variably structured text for each card.

Table 1. Optical character recognition engines tested

Name	**Developer**	**Version**	**Operating System**
ABBYY FineReader PDF	ABBYY	15	Windows
Acrobat Pro 64-bit	Adobe	2022	Windows
Tesseract in LayoutParser	Open Source	5.2	Linux
OmniPage Professional	Nuance (Kofax)	18	Windows
Tesseract in Python-tesseract	Open Source	5.0	Windows

Next, we trained a layout model to perform *Document Image Analysis (DIA)* using the LayoutParser library (Shen et al., 2021). We selected PrimaLayout, a layout analysis model trained on a dataset of magazines, and technical and scientific publications, as our base (Antonacopoulos et al., 2009). Our goals for training a custom model were to: 1) recognize distinct blocks of text from index cards as distinct text fields; and 2) classify each text block with its corresponding field label based on its position.

We used the truth deck we created in LabelStudio to train a custom layout detection model. We split the truth deck into 80% training (400 cards) and 20% testing (100 cards) and updated the PrimaLayout model. To customize our layout detection model with our training data, we used the Fast R-CNN implementation (Girshick, 2015) available in Detectron2, a computer vision library for object detection and image segmentation. We trained each model version using a computing node with a graphics processing unit (GPU) on Great Lakes, a high-performance computing platform available for research use through the University of Michigan. It took approximately fourteen hours to process all 25,744 index cards with our final DIA model.

Finally, we compared three approaches for *parsing* the digitized text to an analysis-ready format. The goal of this step was to capture the original structure of the source information (i.e., card layout) in a structured, tabular output file. The three methods for structuring the output that we compared were: 1) regular expressions (*RegEx*) applied to the OCR text output; 2) document image analysis (DIA) text bounding box delineation; and 3) named entity recognition (NER) classification of unstructured text.

First, we crafted RegEx to segment the OCR output. We took advantage of recurring patterns, such as mortgagor names' occurring on the first line and comma delimiting of city/state pairs, to structure the output. Second, we relied on Tesseract, an open-source OCR engine, to extract text within each bounding box predicted by our custom DIA model. This allowed us to associate the label of each bounding box with the extracted text and generate a structured output. Third, we used spaCy, a natural language processing (NLP) library (Montani et al., 2020), to predict entity types in the unstructured OCR text output. Using the labels available in the pre-trained English pipeline available in spaCy, we created mappings between the following entity types and fields we defined for the index cards: 1) agency: organization (ORG); 2) amount: money (MONEY); 3) location: countries, cities, states (GPE); 4) name: person

(PERSON); and 5) mortgage id: product (PRODUCT). We report the results from these three parsing approaches in **Table 4** in the Results section.

Results

OCR accuracy

We evaluated the output of each of the five OCR methods tested using two measures. To calculate these, we used the PRImA Text Evaluation Tool (OCR Performance Evaluation) 1.5 from Pattern Recognition & Image Analysis (PRImA) Research Lab (PRImA Research Lab, 2018). First, we looked at character accuracy (the percent of characters not needing to be changed to align with ground truth text), which is one of the most commonly-used measures used to evaluate OCR accuracy, along with its inverse, character error rate (Neudecker et al., 2021, Rice, 1996). Second, because of the semi-structured nature of the mortgage records and the relatively low importance of word order from line to line (e.g., it does not matter what order the output text shows the mortgagor's city relative to the mortgage amount), we also measured flex character accuracy (Clausner et al., 2020), which is well-suited to measure OCR accuracy for materials with complex layouts or with content whose word order is less of a priority than the accuracy within individual words. To average the accuracy across the 100-card sample, we weighted each card based on the number of characters on the card.

Table 2 presents the results of these two measures, both of which range from zero (OCR output differs completely from the truth data) to one (OCR output perfectly matches the truth data). For example, ABBYY FineReader's character accuracy of 0.84 means that 84% of characters do not need to be changed for the OCR to align with the ground truth text, conforming to the word order of the ground truth text. The flex character rate of 0.98 indicates that 98% of characters do not need to be changed, when word order is not a consideration. We found that Tesseract in a pytesseract workflow was the most accurate method of OCR, by both measures, with nearly perfect flex character accuracy. However, ABBYY Finereader and OmniPage both performed very well also, especially using the flex character accuracy measure.

Table 2. Character Accuracy and Flex Character Accuracy for each OCR method

OCR method	Character Accuracy	Flex Character Accuracy
ABBYY FineReader	0.84	0.98
Adobe Acrobat	0.80	0.84
Tesseract in LayoutParser	0.83	0.87
OmniPage	0.90	0.95
Tesseract in pytesseract	0.95	0.99

Performance metrics

We also evaluated the performance of our custom layout detection model in the DIA workflow. **Table 3** summarizes average precision (AP) scores for each text field category and for our best model overall. For example, AP of 88.9 for the text field category of Agency indicates an overlap of 88.9% between ground truth and predicted bounding boxes. Percent overlap is often compared to a threshold value, which is used to distinguish false from true positives. We evaluated model performance based on the overlap (i.e., Intersection over Union, or IoU) between ground truth bounding boxes provided in the held-out test set and bounding boxes predicted by the model. If the overlap exceeded a threshold value (in our case, 50%), the model prediction was considered correct (R. Padilla et al., 2020). We used this metric to determine how reliably the model's object detection corresponded to that of a human annotator. We trained our model iteratively until we saw only marginal gains in precision (i.e., correct positive predictions) per label category. We were satisfied with our model's ability to identify bounding boxes containing text and correctly label the text region. For example, our final layout model reliably draws a bounding box around the name of a mortgagor on a scanned index card and classifies the text that it contains as a "Name".

Table 3. Average precision (AP) per text field category

Agency	**Amount**	**ID**	**Location**	**Name**	**Other**	**Status**	***Overall***
88.9	71.5	71.0	70.1	66.2	50.2	79.6	71.1

Parsing accuracy

After finalizing the custom layout detection model, we compared text parsing performance using regular expressions (RegEx), document image analysis (DIA) with LayoutParser, and named entity recognition (NER) with spaCy. Initially, we only parsed the data into five fields due to limitations in the pre-trained entities in our NER model. We used our hand-keyed truth deck to calculate the character error rate (CER) for each approach[1]. **Table 4** summarizes the CER for each method, which indicates the percentage of characters in the parsed output that differ from the truth data, with zero meaning no difference between the OCR output and the truth data (Neudecker et al., 2021). Overall, across all text field categories, we found that text extraction with Tesseract (OCR) and LayoutParser (DIA) was superior to spaCy (NER) and outperformed the use of RegEx for several categories, such as "Name" and "Agency".

Both DIA and RegEx parsed the OCR data much more accurately than the NER model, and both were able to parse the text into much more granular categories as well. Given that parsing with RegEx performed better on some text fields, such as "Location", we also experimented with combinations of approaches. We used RegEx to extract cities and states from the "Location" and "Agency" fields. We also split "Name" into first and last names, and if a second person was named, we parsed the name into a separate field. We ultimately compared DIA and RegEx parsing on 11 fields.

[1] We calculated CER using python code based heavily on xer (Puigcerver, 2014).

Table 4. Character error rate (CER) for each parsing approach per text field category

Field	Regular Expressions (RegEx)	Document Image Analysis (DIA)	Named Entity Recognition (NER)
Agency	4.41	1.73	56.45
Location	8.98	17.23	35.99
Amount	20.96	33.80	53.66
Name	13.73	5.38	73.37
ID	37.91	15.59	98.50
Overall	10.59	9.49	57.02

Table 5 presents the character error rates measured for these parsing methods. We found that post-processing DIA text output using RegEx resulted in a structured text output with the lowest CER overall. At the field level, each parsing method had strengths and weaknesses, with DIA outperforming RegEx in six of 11 fields. The final workflow uses DIA and OCR to extract text from bounding boxes and post-process the output with RegEx to parse it into a tabular file with 12 fields (the 11 listed in **Table 5** plus the name of the scanned card's image file name). **Table 6** in **Appendix A (Supplementary File 1)** provides an example of the output file generated from our index digitization and parsing workflow.

Table 5. Character error rate (CER) for Regular Expressions (RegEx) and Document Image Analysis (DIA) approaches per text field category–more granular parsing

Field	RegEx alone	DIA (LayoutParser) with RegEx post-processing
Last name	0.54	4.35
Person 1 name	14.52	6.93
Person 2 name	24.68	4.75
Amount	20.96	33.80
ID	42.14	15.59
City	1.58	24.09
State	16.67	37.06
Agency	4.41	1.73
Agency city	11.53	3.44
Agency state	11.11	10.88

Status	1.95	4.75
Overall	8.84	8.48

Discussion

This paper develops a workflow for digitizing and parsing text from historical paper-based collections, such as index cards. While technologies such as OCR have been available for several decades and make it possible to extract high-quality text from scanned images, off-the-shelf OCR does not yet take advantage of layout information to structure output (Neudecker et al., 2021). Omitting layout contexts from the processing of paper-based collections results in flat, unstructured text. Even if the OCR output text is high quality, it still requires substantial manual processing to delineate separate fields for record linkage and analysis, which are necessary precursors for research use (Stančić & Trbušić, 2020). Manual processes are often the main bottlenecks in digitization workflows (Blanke et al., 2012). Hand transcription, even with the aid of semi-automated tools, limits the volume of data that can be processed and poses the risk of introducing additional sources of human error.

We have incorporated layout detection methods to digitize and parse historical records. The workflow we proposed combines off-the-shelf, state-of-the-art OCR technology with a custom Document Image Analysis (DIA) model to create a high-quality, structured text dataset for historical research. Our workflow trains a DIA model without much additional overhead, making it possible for archives to implement it for other index card digitization and related projects. The main requirements for training our DIA model were: 1) the creation of truth decks for validating and training; and 2) access to a high-performance computing environment for training and updating the model. The creation of truth decks was streamlined through the use of existing open-source tools such as LabelStudio. We were also able to access a high-performance computing environment through the University of Michigan. Many academic institutions provide reduced-cost HPC environments for use in academic projects.

Taking advantage of document layout brings computational approaches into closer alignment with human judgments and processes. Computer vision research, in particular, seeks to enable computers to derive information from images and other visual inputs (Zhong et al., 2019). For example, deep learning is already making it possible for models to "learn" the layout of a given document and perform various tasks, such as image segmentation and text extraction (Shen et al., 2021). Incorporating deep learning models into existing records management and digitization efforts in archives holds high potential. The NARA catalog (*National Archives and Records Administration*, n.d.) has over 6200 index card-based series that may be useful for researchers if they were processed using our digitization (see **Table 7** in **Appendix B (Supplementary File 2)** for a few relevant examples. Other archives, libraries, and agencies likely have thousands of other similar collections.

The workflow we present brings several advantages to researchers and practitioners. For one, adoption may improve efficiency, freeing up human expertise for other valuable curation tasks, like quality checks and metadata creation, which ensure the discoverability and usability of digital collections. In addition, workflows that leverage deep learning may also support human curators in preparing “collections as data” by making implicit context, like layout information, structurally explicit (T. Padilla et al., 2019). For example, in newspaper digitization efforts, the inclusion of layout information and identification of linked entities supports large-scale network analysis and pattern detection (Lee et al., 2020). Making implicit information (e.g., layout information) explicit increases the analytical utility of the data product for a wider range of scholars and computational research methods.

Conclusion

Paper-based historical records contain rich layout information that must be incorporated to effectively digitize and parse these records into analyzable data. Through the use of deep learning with Document Image Analysis (DIA), we automatically recovered information about document layouts. We applied this technique to process a collection of administrative records related to the Servicemen’s Readjustment Act of 1944, also known as the G.I. Bill. We showed how to use DIA in combination with standard OCR and RegEx approaches to extract high-quality, structured text from scanned images. In summary, this paper contributes: 1) a workflow using scanning, optical character recognition, and deep learning to digitize and parse index cards; and 2) a novel, analysis-ready dataset for historical research. The workflow demonstrates the feasibility of incorporating deep learning into archives’ existing digitization and parsing efforts. In addition, the digitized dataset is ready to be linked with additional data sources to further increase its analytical research utility.

Acknowledgments

This work was supported by a Propelling Original Science (PODS) grant, “Images to Integrated Data”, provided by the Michigan Institute for Data Science (MIDAS) at the University of Michigan. We are grateful to the Digitization Division staff at the Office of Research Services at the National Archives and Records Administration–and specifically to Denise Henderson–for their guidance and collaboration in the scanning of the paper documents. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

Appendix A (Supplementary File 1): Example Output

Table 6. Example of document image analysis layout model text output post-processed with regular expressions

Image	City	State	Last name	Person 1 name	Person 2 Name	Amount	Agency	Agency City	Agency State	ID	Status
783095-01-0003.jpg	Detroit	Mich	Abbenga	Arnold N	Geraldine	$4,100,	Detroit Agency - Agent's Serial No. PC 603, James T, Barnes & Company,	Detroit	Mich	LH-31228	(V4) FIRST MORTGAGE LOAN GUARANTEED
783095-01-0004.jpg	Tulsa	Oklahoma	Abbey	Leonard Ray	Barbara J oan	$8,000.	Agent's Serial No. 44, W. R. Johnston & Co., Inc.,	Oklahoma City	Oklahoma	LH-5746	(Va) FIRST MORTGAGE LOAN GUARANTEED

Appendix B (Supplementary File 2): Index Card Collections

Table 7. Examples of index card collections listed in the National Archives and Records Administration (NARA) Catalog–all information copied directly from NARA Catalog website (National Archives and Records Administration, n.d.)

Title	NARA ID	Description
Card Index of Licensees, ca. 1918–ca. 1918	5111291	This series consists of a card index listing businesses (sometimes listed under the name of an individual, otherwise listed by the name of the firm) licensed by the Iowa State Food Administration. Both wholesalers and retailers are included. The name and address of the establishment is listed, along with an unidentified sequence of numbers, the last of which is the number of the form on which the business was obliged to report to the United States Food Administration.
Awards Files Card Index, 1944–1945	611132	This series consists of an index of the awards and decorations given to soldiers of the 44th Infantry Division during the campaigns in northern France, the Rhineland, and central Europe during World War II. The awards referenced in this series include the Silver Star, Bronze Star, Air Medal, and Purple Heart. Each index entry provides the name of the soldier given the award, the type of award given, and the date of the award.
Publications Card Files, 1944–1945	624400	This series consists of cards showing the name of a propaganda publication or leaflet; the date of publication; the publisher; the number of copies printed; the date of pickup for distribution; and the date, number of copies, and location of dissemination. Many of the cards have a copy of the reference publication attached.
Card Index of Trusts, 1917–1934	6879969	This series consists of an index for trusts established for individuals or companies whose property was seized by the Alien Property Custodian. The cards in this series include spaces for name and address, trust number, date opened, report number, ticket number, and other information. Sometimes, only a name and trust number are included on the card.
Death Certificate Card Index, 1914–February 1915	7408557	The series contains an index that records deaths that occurred in the Canal Zone. The records include information concerning the deceased such as name, age, color, sex, nationality, occupation and employment, residence, address, nature of illness and cause of death, attending physician, date of death, and grave number.

References

ABBYY. (2019). *ABBYY FineReader PDF* (Version 15) [Computer software]. https://pdf.abbyy.com/media/1676/users_guide.pdf

Adobe. (2022). *Acrobat Pro 64-bit* (Version 2022) [Computer software]. https://www.adobe.com/acrobat/acrobat-pro.html

Amujala, S., Vossmeyer, A., & Das, S. R. (2023). Digitization and data frames for card index records. *Explorations in Economic History*, *87*, 101469. https://doi.org/10.1016/j.eeh.2022.101469

Antonacopoulos, A., Bridson, D., Papadopoulos, C., & Pletschacher, S. (2009). A Realistic Dataset for Performance Evaluation of Document Layout Analysis. *2009 10th International Conference on Document Analysis and Recognition*, 296–300. https://doi.org/10.1109/ICDAR.2009.271

Arthur, K., Byrne, S., Long, E., Montori, C. Q., & Nadler, J. (2004). *Recognizing Digitization as a Preservation Reformatting Method*. *33*(4), 171–180. https://doi.org/10.1515/MFIR.2004.171

Blanke, T., Bryant, M., & Hedges, M. (2012). Open source optical character recognition for historical research. *Journal of Documentation*, *68*(5), 659–683. https://doi.org/10.1108/00220411211256021

Booth, J. M., & Gelb, J. (2006). *Optimizing OCR accuracy on older documents: a study of scan mode, file enhancement, and software products* (v2.0 ; pp. 1–5). U.S. Government Printing Office.

Brahney, K. (2015). *Information Extraction from Semi-Structured Documents MSci. Computer Science with Industrial Experience 05/06/2015*. http://miami-nice.co.uk/information-extraction-from-docs.pdf

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2020). Flexible character accuracy measure for reading-order-independent evaluation. *Pattern Recognition Letters*, *131*, 390–397. https://doi.org/10.1016/j.patrec.2020.02.003

Federal Agencies Digital Guidelines Initiative. (2022). *Technical guidelines for digitizing cultural heritage materials* (No. 3.5; pp. 73–81). Still Image Working Group. https://www.digitizationguidelines.gov/guidelines/digitize-technical.html

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448. http://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

Hoffstaetter, S. (2021). *Python-tesseract* (Version 0.3.8) [Computer software]. https://github.com/madmaze/pytesseract

Index to Loans on Veterans Administration Guaranteed Mortgages, 1946 – 1954. (n.d.). [Data set]. In *National Archives NextGen catalog*. https://catalog.archives.gov/id/783095

Irmak, U., & Kraft, R. (2010). A scalable machine-learning approach for semi-structured named entity recognition. *Proceedings of the 19th International Conference on World Wide Web*, 461–470. https://doi.org/10.1145/1772690.1772738

Kasturi, R., O'Gorman, L., & Govindaraju, V. (2002). Document image analysis: A primer. *Sadhana*, *27*(1), 3–22. https://doi.org/10.1007/bf02703309

Katznelson, I., & Mettler, S. (2008). On race and policy history: A dialogue about the G.I. bill. *Perspectives on Politics*, *6*(3), 519–537. https://doi.org/10.1017/s1537592708081267

Lee, B. C. G., Mears, J., Jakeway, E., Ferriter, M., Adams, C., Yarasavage, N., Thomas, D., Zwaard, K., & Weld, D. S. (2020). The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3055–3062. https://doi.org/10.1145/3340531.3412767

Leetaru, K. (2008). Mass book digitization: The deeper story of Google Books and the Open Content Alliance. *First Monday*, *13*(10). https://doi.org/10.5210/fm.v13i10.2101

Leimer, J. (1962). Design factors in the development of an optical character recognition machine. *IRE Transactions on Information Theory*, *8*(2), 167–171. https://doi.org/10.1109/TIT.1962.1057696

Lischer-Katz, Z. (2022). The emergence of digital reformatting in the history of preservation knowledge: 1823–2015. *Journal of Documentation*, *78*(6), 1249–1277. https://doi.org/10.1108/JD-04-2021-0080

Lu, C., Bing, L., Lam, W., Chan, K. I., & Gu, Y. (2013). Web entity detection for semi-structured text data records with unlabeled data. *International Journal Of. Computational Linguistics and Applications*, *4*(2), 135–150. http://www.ijcla.org/2013-2/IJCLA-2013-2-pp-135-150-Web.pdf

Montani, I., Honnibal, M., Honnibal, M., Van Landeghem, S., Boyd, A., Peters, H., Samsonov, M., Geovedi, J., Regan, J., Orosz, G., McCann, P. O., Kristiansen, S. L., Altinok, D., Roman, Fiedler, L., Howard, G., Bozek, S., Phatthiyaphaibun, W., Amery, M., … Patel, A. (2020). *spaCy: Industrial-strength natural language processing in Python* (Version v3) [Computer software]. https://doi.org/10.5281/zenodo.1212303

Nagy, G. (1992). At the frontiers of OCR. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, *80*(7), 1093–1100. https://doi.org/10.1109/5.156472

National Archives and Records Administration. (n.d.). National Archives Catalog. https://catalog.archives.gov/

Navarrete, T., & Owen, J. M. (2011). Museum libraries: how digitization can enhance the value of the museum. *Palabra Clave (La Plata)*, *1*(1), 12–20. http://www.scielo.org.ar/img/revistas/pacla/v1n1/html/v1n1a03.htm

Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., & Pletschacher, S. (2021). A survey of OCR evaluation tools and metrics. *The 6th International Workshop on Historical Document Imaging and Processing*, 13–18. https://doi.org/10.1145/3476887.3476888

Nuance Communications, Inc. (2011). *OmniPage Professional* (Version 18) [Computer software]. Nuance Communications, Inc.

Packer, T. L. (2011). Performing information extraction to improve OCR error detection in semi-structured historical documents. *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, 67–74. https://doi.org/10.1145/2037342.2037354

Padilla, R., Netto, S. L., & da Silva, E. A. B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

Padilla, T., Allen, L., Frost, H., Potvin, S., Roke, E. R., & Varner, S. (2019). *Always Already Computational: Collections as data: Final report*. https://digitalcommons.unl.edu/scholcom/181/

Pondenkandath, V., Seuret, M., Ingold, R., Afzal, M. Z., & Liwicki, M. (2017). Exploiting State-of-the-Art Deep Learning Methods for Document Image Analysis. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, *05*, 30–35. https://doi.org/10.1109/ICDAR.2017.325

PRImA Research Lab. (2018). *PRImA Text Evaluation Tool* (Version 1.5) [Computer software]. https://www.primaresearch.org/tools/PerformanceEvaluation

Puglia, S. T., Reed, J., & Rhodes, E. (2005). *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images*. National Archives and Records Administration. https://play.google.com/store/books/details?id=M41NLKdXIdkC

Puigcerver, J. (2014). *xer*. https://github.com/jpuigcerver/xer

Rice, S. V. (1996). *Measuring the accuracy of page-reading systems* (T. A. Nartker (Ed.)) [University of Nevada, Las Vegas]. https://www.proquest.com/dissertations-theses/measuring-accuracy-page-reading-systems/docview/304329395/se-2

Servicemen's Readjustment Act of 1944, 78th Congress, Pub. L. 346, 18 (1944). https://hdl-handle-net.proxy.lib.umich.edu/2027/umn.31951d03569283l

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *Document Analysis and Recognition – ICDAR 2021*, 131–146. https://doi.org/10.1007/978-3-030-86549-8_9

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, *2*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Stančić, H., & Trbušić, Ž. (2020). Optimisation of archival processes involving digitisation of typewritten documents. *Aslib Journal of Information Management*, *72*(4), 545–559. https://doi.org/10.1108/AJIM-11-2019-0326

Stevens, M. E. (1961). *Automatic Character Recognition: A State-of-the-Art Report*. U.S. Department of Commerce. https://hdl-handle-net.proxy.lib.umich.edu/2027/mdp.39015077289836

Tesseract. (2021). *Tesseract OCR* (Version 5.0) [Computer software]. https://github.com/tesseract-ocr/tesseract

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., & Liubimov, N. (2020). *LabelStudio:Data labeling software* (Version 1.7) [Computer software]. https://github.com/heartexlabs/label-studio

United States Department of Veterans Affairs. (2013). *History and Timeline—Education and Training*. https://www.va.gov/education/about-gi-bill-benefits/

Zaid, C. (1973). *Preliminary Inventory of the Records of the Reconstruction Finance Corporation, 1932-1964 (Record Group 234)*. National Archives & Records Service. https://hdl-handle-net.proxy.lib.umich.edu/2027/uiug.30112101560024

Zhong, X., Tang, J., & Jimeno Yepes, A. (2019). PubLayNet: Largest Dataset Ever for Document Layout Analysis. *2019 International Conference on Document Analysis and Recognition*

(ICDAR), 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166