

Beyond the Bounding Box

Entity Extraction and the Remediation of Newspaper Directories

Daniel Evans, *University of Illinois; Urbana–Champaign*

DOI: 10.55520/TBA

Introduction

Business directories are often treated as static sources of information. Published annually, their contents shifted from edition to edition as publishers appeared and vanished, and as newspaper titles and addresses changed. Despite this fluidity, scholars have largely treated them as static reference works rather than dynamic records of a changing media landscape. This article argues that extracting information from nineteenth- and twentieth-century business directories at scale can serve as a critical site for historical recovery, particularly for serial publications that have resisted systematic study.

To make this argument, I show how these directories can be remediated, or transformed from fixed archival pages into structured, analyzable datasets composed of traceable entities. A single directory entry, for instance, might yield a publisher name, associated street address, publication frequency, and circulation figures. Linked across editions and cross-referenced with other records, these discrete data points become evidence of patterns that no individual page could reveal on its own. Examples of these patterns include the geographic clustering of ethnic newspapers, the lifespan of religious periodicals, and the market conditions shaping who got listed and who did not. In this sense, the decisive metric for the sparse text pulled from these pages is not the fidelity of optical character recognition (OCR) but rather the ability to consistently identify the entities that make such linkage possible.

Extracting structured data from historical business directories at scale has become feasible only recently. For much of the twentieth and early twenty-first centuries, methods such as the Text Encoding Initiative (TEI), early OCR technologies, or model-specific fine-tuning defined the scale and scope of what researchers could do with digitized materials. Today, advances in transformer-based computer vision models (deep learning architectures that process input data in parallel and learn contextual relationships between elements, enabling them to interpret

both text and images) have made it practical to process historical documents that once demanded painstaking manual transcription or highly specialized annotation. Yet efficiency alone is not the most important outcome. As Jessica Marie Johnson and Jordan Taylor have shown, archival absence is never neutral; the silences in Black and abolitionist print cultures are often the result of deliberate minimization, market forces, or structural inequities.¹ I frame these remediating outcomes as successful if they surface traces of what has been excluded, distorted, or rendered invisible, rather than simply recovering text more accurately.

To do so, I show how these directories can be remediated as dynamic, analyzable datasets, expanding their interpretive horizons. In particular, I argue that the decisive metric for these directories is both OCR fidelity and the consistent extraction of entities, such as the publishers, locations, frequencies, and circulation data, that transform the record into structured evidence. This extraction enables large-scale data analysis while also creating new opportunities for editors to contextualize and link serial publications that have historically resisted systematic study.

I take as a case study the newspaper directories of two long-running American advertising agencies: George P. Rowell and N. W. Ayer & Son.² Both firms published directories during the second half of the nineteenth century through the twentieth century. Both are readily available as digitized objects. While the development of the directory should be read within the context of the greater nineteenth-century shift toward tabulation, new advancements in machine learning allow contemporary scholars to computationally analyze this form of literature. This makes them well suited for the kind of large-scale computational analysis I describe in this paper. They also exemplify an informational form that humanities scholars have largely overlooked. As for-profit enterprises, these agencies transformed scattered publication details into standardized, tabular listings designed to help advertisers buy space in newspapers. Their directories were structured for commercial gain and were not designed with the intention of cultural preservation. I argue that this commercial objective shaped what was recorded, what was omitted, and how entries were organized. Reading these directories as data sources rather than neutral reference works means reckoning with the choices these agencies made.

I.

A LIST OF THE NEWSPAPERS AND OTHER PERIODICALS IN THE UNITED STATES AND TERRITORIES, AND THE DOMINION OF CANADA AND BRITISH COLONIES; ARRANGED ALPHABETICALLY BY TOWNS, GIVING NAME, DAYS OF ISSUE, POLITICS OR GENERAL CHARACTER, FORM, SIZE, SUBSCRIPTION PRICE PER YEAR, DATE OF ESTABLISHMENT, EDITOR'S AND PUBLISHER'S NAMES, CIRCULATION, ETC.

ALABAMA.	ALABAMA.
ABBEVILLE, Henry Co. Register; Saturdays; democratic; four pages; size 22x32; subscription \$3; established 1866; Kelly & Quillin, editors and publishers; claims 500 circulation. ATHENS Post; Saturdays; independent; four pages; size 24x36; subscription \$3; established 1865; T. B. Reynolds, editor and publisher; claims 1,500 circulation. BUTLER, Choctaw Herald; Wednesdays; democratic; four pages; size 21x28; subscription \$3; established 1868; M. L. Yeatman, editor and publisher. CAMDEN, Wilcox News and Pacificator; Tuesdays; democratic; four pages; size 22x32; subscription \$3; established 1866; W. B. H. Howard, editor; F. P. M. Gilbert, proprietor. CAMDEN, Wilcox Vindicator; Fridays; democratic; four pages; size 22x32; subscription \$3; established 1867; C. L. Scott, editor and publisher. CARROLLTON, West Alabamian; Wednesdays; democratic; four pages; size 24x34; subscription \$3; established 1849; Henry & Gilbert, editors and publishers; claims over 600 circulation. CENTRE, Cherokee Advertiser; Thursdays; democratic; four pages; size 22x32; subscription \$2 50; established 1866; John M. Cothran, publisher; Mrs. Jane H. Stiff, proprietor. CHAMBERS Tribune; (no report.) CLAIBORNE, Monroe Eagle; Wednesdays; democratic; four pages; size 28x42; subscription \$3; established 1868; Rev. W. W. Spence, editor; S. B. Baird, publisher; claims 800 circulation. CLAYTON Banner; Thursdays; democratic; four pages; size 21x28; subscription \$3; established 1851; John Post, editor and publisher; claims 500 circulation. COLUMBIANA, Shelby Co. Guide; Thursdays; democratic; four pages; size 22x32; subscription \$2 50; established 1868; W. Roberts, editor and publisher; claims 500 circulation. DADEVILLE, Tallapoosa News; Fridays; republican; four pages; size 22x32; subscription \$2; established 1866; J. W. Johnston, editor and publisher; claims 1,100 circulation. DECATUR, Alabama Republican; Wednesdays; republican; four pages; size 25x38; subscription \$2 50; established 1867; C. C. Sheats, editor and publisher. DEMOPOLIS, New Era; Wednesdays; democratic; four pages; size 22x32; subscription \$4; established 1865; H. E. & F. A. Monroe, editors and publishers; claims 500 circulation. ELYTON Herald; Mondays; democratic; four pages; size 21x34; subscription \$2 50; established 1865; H. A. Hale, editor and publisher. EUFAULA News; tri-weekly; Tuesdays, Thursdays and Saturdays; democratic; four pages; size 23x32; subscription \$9; established 1866; John Black, editor and publisher.	EUTAW Whig and Observer; Thursdays; democratic; four pages; size 24x36; subscription \$4; established 1841; W. O. Monroe, editor and proprietor; claims 600 to 700 circulation. FAYETTE Watchman; Saturdays; democratic; four pages; size 23x33; subscription \$2; established 1852; James S. Harrison, editor and publisher; John C. Robertson, proprietor; claims 700 circulation. FLORENCE Journal; Thursdays; four pages; size 24x34; subscription \$3; David R. Lindsay, editor and publisher. FLORENCE, Literary Index; Thursdays; independent; four pages; subscription \$4; established 1865; Isaac S. Barr, editor and publisher; claims 300 circulation. GADSDEN, Republican Union; Thursdays; four pages; size 21x28; subscription \$2; established 1869; P. J. Smith, editor and publisher. GADSDEN Times; Fridays; democratic; four pages; size 22x32; subscription \$2; established 1867; L. W. Grant, editor and publisher; claims 500 circulation. GAINESVILLE, News; Wednesdays; democratic; four pages; size 16x22; subscription \$2; established 1868; J. D. Coward, editor and publisher. GREENSBORO, Alabama Beacon; Saturdays; democratic; four pages; size 26x40; subscription \$4; established 1840; John G. Harvey, editor and proprietor. GREENVILLE Advocate; tri-weekly; Tuesdays, Thursdays and Saturdays; democratic; four pages; size 24x36; subscription \$6; established 1865; James M. Whitehead, editor; Jas. B. Stanley, publisher and proprietor; claims 1,000 to 1,100 circulation. GROVE HILL Herald; (no report.) HAYNEVILLE Chronicle; (no report.) HAYNEVILLE Examiner; Wednesdays; democratic; four pages; size 22x32; subscription \$3; established 1868; Brewer & Michael, editors and publishers. HUNTSVILLE Advocate; semi-weekly; Tuesdays and Fridays; republican; four pages; size 26x36; subscription \$4; established 1865; W. B. Figures, editor and publisher; claims 1,000 circulation. HUNTSVILLE Democrat; every evening except Monday, and Weekly, Fridays; democratic; four pages; size—daily 22x32, weekly 25x38; subscription—daily \$10, weekly \$3; established—daily 1868, weekly 1823; J. Withers Clay, editor and publisher. HUNTSVILLE Independent; every morning except Monday, and Weekly, Saturdays; democratic; four pages; size—daily 22x32, weekly 26x40; subscription—daily \$10, weekly \$3; established—daily 1866, weekly 1855; Dew, Young & Co., editors and publishers.

Figure 1. A page from Rowell's American Newspaper Directory, 1869.

In undertaking this work, my primary aim is to find an afterlife for these directories and to make them available for reading and reuse well beyond their original audience and moment.³ At the same time, as an information scientist, I am interested in what these directories can contribute to newspaper metadata and discoverability more broadly. These goals are linked. The same process that transforms a directory page into a structured dataset also generates evidence that can improve how we identify, locate, and contextualize historical serial publications.

Ultimately, this article contributes both a technical workflow and a methodological claim. By treating directories as tabular data, or information organized in rows and columns where each row represents a single entry and each column a category such as publisher or location, instead of prose, scholars can generate computational spreadsheets from precomputational sources. But the ambition extends beyond data extraction. Following Ryan Cordell's concept of "speculative bibliography," I argue that structured data drawn from these directories can help us imagine what lies beyond the archival record itself.⁴

Clean Data and Data Remediation

Realizing this ambition, however, requires engaging with a set of prior debates regarding how digitized cultural heritage materials become usable data and about the assumptions embedded in that transformation. The following section situates this project within scholarship on data remediation, OCR quality, and the collections-as-data paradigm before turning to the specific methods I employ.

This work takes the "collections as data" paradigm as a starting point for conceptualizing analog enumerative bibliographies.⁵ This paradigm argues for treating digitized cultural collections both as images and textual objects for reading, and as structured, computable data sources. My work focuses on extracting data from late nineteenth- and early twentieth-century American newspaper directories by using a document image analysis (DIA) approach to conceptualize entire bibliographic documents themselves as tables, instead of simply texts, as I train natural language processing (NLP) and transformer-based machine learning models.⁶

In centering attention on the data that is extracted via optical character recognition from a document, this article highlights why optimizing OCR quality is central to our understanding of a digital text. Far from being a neutral transcription step, OCR has historically mediated what is legible, searchable, and analyzable in digital archives. As other researches have demonstrated, I highlight how "messy" OCR continues to shape downstream processes and interpretations.⁷ By situating these limits alongside recent advances in machine learning, I argue that the role of OCR has shifted from a fragile bottleneck to an enabling condition for both preservation and new forms of research. These shifts also set the stage for moving beyond text to entities, and positions OCR as the hinge for the broader remediation of cultural heritage records.

OCR is a key aspect of turning physical documents into machine-readable text and is often considered a "solved" problem.⁸ It stands at the end of a pipeline that converts scanned documents, images, or microfiche into digitized data. Like its name suggests, OCR recognizes the alphanumeric characters and words within an image, thereby allowing for the extraction of machine-encoded text. This technology allows for the rapid duplication of archival and historical documents, along with enhanced searchability and discoverability when searching across a large corpus.

Because OCR makes documents searchable, there has been a historic reliance on the textual aspects of information retrieval and content mining. The page image is itself “an object to be passed through” on the way to OCR or textual analysis, “rather than studied as something potentially meaningful in itself.”⁹ OCR ensures discoverability of documents by allowing for an increased accessibility of digitized texts over nondigitized ones. Ian Milligan cites the “Matthew’s effect” of digitized resources where “the rich (digitized) got richer in terms of use, whereas the poor (un-digitized) got poorer.” He points to a 991% increase in citations for the *Toronto Star* after it was digitized and OCR’d.¹⁰ This “Matthew’s effect” of digitized resources underscores the importance of digitization of archival documents.

Aside from discoverability, OCR quality has also been proven to affect the success of downstream assessment tasks. Others have developed strategies to improve the quality of OCR text to minimize transcription errors through a process known as post-correction, whether through manual intervention or automated pipelines or, as a more drastic step, to re-OCR the entire corpus themselves.¹¹

My work extends these assessments by first testing an OCR post-correction model to establish a baseline for textual accuracy and then proposing a multimodal model for entity extraction. This shift moves the focus from OCR errors to the entities themselves, from treating records as textual objects to understanding what those objects represent. Ultimately, I argue that these models enable a new form of remediation, in which cultural heritage data is repurposed from physical record into digital text and into structured ontological entities that can be queried, compared, and recontextualized.

The need for an increased focus on data remediation of our cultural heritage records is only heightened in this moment of massive data generation and increased reliance on artificial intelligence (AI). Generative AI systems have aided technologists in this effort by allowing for the rapid development of improved training and language models that produce tools, such as more accurate optical character recognition, automated indexing, and more efficient coding systems. Despite these improvements and increasingly widespread adoption, these technologies are still somewhat opaque to users. Both their training data and model algorithms are often closely guarded proprietary secrets.

Indeed, in this specific moment data collection itself needs to be reexamined with a more critical eye. Large indiscriminate datasets are being used to train large language models (LLMs) and they in turn replicate the structural biases and stereotypes inherent to society. As other researchers have observed, there is space to draw from archival practices and develop procedures and guidelines, especially in NLP and computer vision, that emphasize fine-grained approaches over size and efficiency when collecting data.¹² While recent work has begun to investigate data collection within machine learning domains, digital humanists and information scientists have an opportunity to provide a critical lens through which to examine dataset creation.¹³

The Data

The data for this project consist of forty years of *George P. Rowell’s American Newspaper Directory*, along with forty-five years of *N. W. Ayer & Son’s American Newspaper Annual*. Rowell’s *Directory* ran from 1869 to 1909 and grew from approximately 5,000 entries to over 20,000. Ayer’s *Annual* began later, running from 1880 until 1925.

The data was collected in two phases:

1. Library of Congress Corpus (Green): This dataset consisted of six years from Rowell's and thirty-five years from Ayer. The total data dump consisted of ~45,600 tiffs (Tagged Image File Format), or ~600 scans per year, which were provided by the Library of Congress in 2023 and were scanned in the early 2000s from microfilm. I manually removed front matter, ads, blank pages, maps, and extraneous materials, resulting in a subset of the 45,000 tiffs.
2. HathiTrust/Internet Archive Corpus (Yellow): This dataset consists of thirty years of Rowell's directories (with four missing) and fourteen years from Ayer. These directories are hosted by HathiTrust, Internet Archive, and by the Georgia State Library. Many of these were scanned from the original directories as full color images.

Together the Green and Yellow phases comprise about ~62,000 relevant tiffs. Due to copyright restrictions, the *N. W. Ayer & Son's American Newspaper Annual* is not publicly available after 1925.

Rowell's	Year	Ayer's
LoC	1869	n/a
HT/IA	1870	n/a
HT/IA	1871	n/a
LoC	1872	n/a
LoC	1873	n/a
LoC	1874	n/a
LoC	1875	n/a
HT/IA	1876	n/a
LoC	1877	n/a
HT/IA	1878	n/a
HT/IA	1879	n/a
HT/IA	1880	LoC
HT/IA	1881	LoC
HT/IA	1882	LoC
HT/IA	1883	LoC
HT/IA	1884	LoC
HT/IA	1885	LoC
HT/IA	1886	LoC
HT/IA	1887	HT/IA
n/a	1888	LoC
HT/IA	1889	LoC
n/a	1890	LoC

Rowell's	Year	Ayer's
HT/IA	1891	HT/IA
n/a	1892	HT/IA
HT/IA	1893	LoC
HT/IA	1894	LoC
HT/IA	1895	HT/IA
n/a	1896	HT/IA
HT/IA	1897	LoC
HT/IA	1898	LoC
HT/IA	1899	HT/IA
HT/IA	1900	LoC
HT/IA	1901	LoC
HT/IA	1902	LoC
HT/IA	1903	LoC
n/a	1904	LoC
HT/IA	1905	LoC
HT/IA	1906	LoC
HT/IA	1907	LoC
HT/IA	1908	LoC
HT/IA	1909	LoC
n/a	1910	LoC
n/a	1911	LoC
n/a	1912	LoC
n/a	1913	LoC
n/a	1914	LoC
n/a	1915	LoC
n/a	1916	LoC
n/a	1917	LoC
n/a	1918	LoC
n/a	1919	LoC
n/a	1920	LoC
n/a	1921	HT/IA
n/a	1922	LoC
n/a	1923	HT/IA
n/a	1924	HT/IA

Rowell's	Year	Ayer's
n/a	1925	HT/IA

■ LoC — Library of Congress corpus
 ■ HT/IA — HathiTrust / Internet Archive corpus
 ■ n/a — Not available

Table 1. Processing workflow of directories.

In terms of composition, the directories follow roughly the same pattern throughout. They list such things as location, frequency of circulation, type of newspaper or periodical, year established, circulation numbers, cost, physical size of paper, but also population of county and town. Like others in the nineteenth century, there is little differentiation between newspaper and periodical; accordingly, the publications they enumerate are best understood under the broader category of serials.¹⁴

From the Rowell and Ayer corpora, I identify and release as an open dataset the African American titles that appear within these directories. Extracted using the same entity recognition workflow, this subset illustrates how directories, originally compiled for commercial advertising purposes, nonetheless contain valuable records of marginalized print cultures. By structuring this data and making it openly available, I aim to provide scholars with a resource that supplements existing bibliographies and underscores the speculative potential of archival recovery. In this sense, the dataset demonstrates how computational remediation can surface titles that were always present in the historical record but have remained difficult to access, contextualize, or analyze at scale.¹⁵

The Method

I test three methods for OCR extraction and cleaning: OCR post-correction, cosine-similarity with transformers, and a hybrid document-processing model. These are ByT5, LLaMA 2, LLaMA 3, and a cloud-based service provided by Google, Document AI, which uses a fine-tuned Gemini 1.5 model. As an initial experiment, I implemented an OCR post-correction approach on a small subset of the Rowell corpus. To establish a reference standard, I utilized hand-keyed Text Encoding Initiative (TEI) data derived from the 1869 edition of *George P. Rowell’s American Newspaper Directory*. This curated TEI dataset functioned as a gold-standard benchmark against which model performance could be evaluated.

Using this gold-standard corpus, I fine-tuned a ByT5 transformer model for the task of OCR error correction. A transformer model is a machine learning model that underlies most current NLP and computer vision systems. The objective of this stage was twofold: first, to assess the feasibility of adapting ByT5 for historical OCR post-correction on noisy digitized data, and second, to generate a model capable of being applied more broadly across the remaining editions of the Rowell directories. In theory, this method could then be extended to the Ayer’s data through the manual hand-keying of a small subset of training data. Although differing in form, the directories’ OCR quality, broadly speaking, is affected by small type, dense columns, alphanumeric codes, symbols, and other artifacts inherited because many of them were digitized microfilm.

When successful, these methods reduce both character error rate (CER) and word error rate (WER) while improving the accuracy of downstream NLP tasks, such as named entity recognition and topic modeling (van Strien). Character or word error rates are calculated by summing the substitutions (S), deletions (D), and insertions

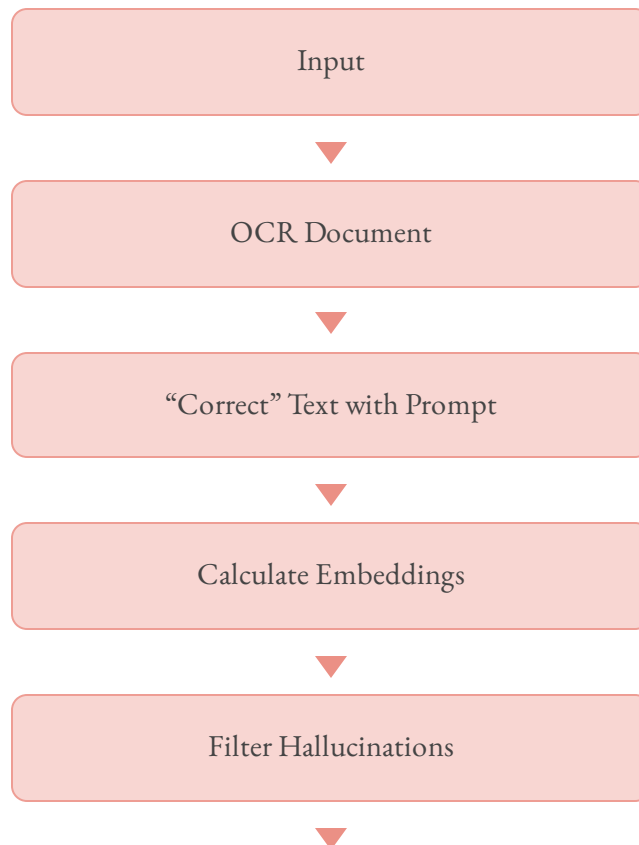
(I) of a given set of characters or words and dividing them by the total number of characters or words (N). They measure the similarity of characters or words across two given texts. The goal in testing various large language models and a multimodal model in this study is to evaluate their ability to accurately extract OCR content from nonstandardized and historical tables.

$$(\text{Character/Word}) \text{ Error Rate} = \frac{S + D + I}{N}$$

S = Substitutions **D** = Deletions **I** = Insertions **N** = Total characters or words

Figure 2. Calculating character and word error rate.

In addition to the ByT5 post-correction model, I implemented a large language model approach to OCR correction. In this method, the model was supplied with a passage of “dirty” OCR output and prompted to generate a corrected version of the text. To mitigate hallucinations, outputs were filtered using cosine similarity between word embeddings. Hallucinations are instances in which the model generates plausible but fabricated text that does not appear in the source document. Cosine similarity is a useful way to gauge the number of hallucinations. In essence, it measures how similar two pieces of text are, using their position in a mathematical vector space. While this strategy has demonstrated success for full-text documents, the sparsity and irregularity of tabular data significantly increase the likelihood of hallucinations. To assess accuracy, the corrected outputs were evaluated against a hand-keyed gold standard using character error rate (CER) and word error rate (WER) metrics.



Output

Figure 3. OCR post-correction with large language models (LLMs).

The uncorrected Tesseract model resulted in a WER of 58.17% and a CER of 20.21%. Among the post-correction methods tested, the ByT5 model produced the most reliable results. It resulted in a WER of 6.49% and a CER of 7.32%. In contrast, both large language models (LLMs) exhibited notable limitations, particularly in the form of hallucinated outputs and constraints imposed by restricted context windows. These issues significantly undermined their performance in comparison to the more targeted post-correction approach. Llama2:7b actually increased its WER and CER to 80.90% and 68.80%, respectively, while Llama3:8b performed marginally better than uncorrected Tesseract, reducing the error rate to 50.40% and 18.62%.

Method	Model	WER	CER
	Uncorrected Tesseract	58.17%	20.21%
Method 1	Byt5	6.49%	7.32%
Method 2	Llama2:7b	80.90%	68.80%
Method 2	Llama3:8b	50.40%	18.62%

Table 2: OCR post-correction success metrics.

In order to be implemented at scale, the Byt5 model would require significantly more training data. It should also be noted that while the Byt5 model performed best in considering WER and CER of tabular data, Llama2 and Llama3 could be useful with natural, or less sparse, text.

Because the ultimate goal of this project is to construct a model capable of reliably producing computational spreadsheets from precomputational tabular data, I reframed the problem as one of recognizing entities rather than optical characters. I chose this approach because the tabular layout of the page meant that the location of each entity could be consistently predicted and I did not need to rely solely on textual extraction. In other words, I was refining data from structured fields rather than running prose and I needed the best entity extraction model to do so. With this orientation, I sought a baseline model that would maximize the accuracy of extracting the categories of data most relevant to my research. After evaluating available options, I selected Google’s Gemini 1.5 model and fine-tuned it through their Google Cloud platform.

By fine-tuning this model, I was not extracting OCR text. Instead, I defined a custom set of entities aligned with the structure of the directories. During annotation, I repeatedly identified features of interest, such as locations, publishers, and genres, and trained the model to recognize these attributes as discrete entities within the data. This

approach enabled the model to move beyond text recognition toward structured data extraction tailored to the requirements of cultural heritage analysis. These entities roughly fit the classifications defined by the directory compilers. The custom entities can be seen in Table 3. Many of these entities and the reasoning behind them are explained in the following sections.

I annotated two hundred documents over the course of two months in the spring of 2025 (70:30 split) to provide a gold standard for fine-tuning. Two hundred was an arbitrary cutoff. These documents were randomly selected but evenly represented both Rowell’s and Ayer’s directories. After fine-tuning, the model reached an F1 score of 0.919 with a precision of 0.922 and a recall of 0.916, indicating high reliability in recognizing and extracting named entities from the directories.

Label	F1 Score
All labels	0.919
newspaper-title	0.918
circulation	0.907
date_established	0.968
genre_or_type	0.949
language	0.952
length	0.903
location	0.921
notes	0.595
publisher_or_editor	0.918
publishing_frequency	0.958
size	0.886
state	0.86
subscription_cost	0.948

Label	F1 Score
title	0.892
year	0.98

Table 3: Entity extraction results. Note that “newspaper_title” refers to a single newspaper title (i.e., *New York Times*), whereas “title” refers to how well the model is able to identify the entire entry (i.e., newspaper_title, language, genre, etc.).

Because the directories were produced annually over decades, OCR errors are partially “absorbed” or mitigated across the dataset. In other words, what is missed in one year might appear in the next. Essentially the redundancy of the directories’ form ensures that entries are not often lost; the accuracy of identifying entities matters more than the accuracy of single line transcriptions.

While the model had a relatively high F1 score, it did not perform consistently across all fields. It struggled, for example, to identify the notes entity. This is acceptable given the high variability of the field and the fact that it was relatively rare compared to the other entities. Other low-scoring entities, such as state, only appear on every other page and can be inferred in subsequent cleanup.

Toward a Speculative Bibliography

While much of the work on OCR in cultural heritage collections has treated accuracy as the primary benchmark of success, the precomputational layout of nineteenth-century business directories complicates this assumption. My early efforts sought to maximize conventional measures such as character error rate (CER) and word error rate (WER), but it became increasingly clear that these metrics, while important, do not fully capture the most valuable evidence in these materials. Because the directories are structured as complex tables, they often resist line-by-line transcription. The decisive question is not how closely the text mirrors the original, but whether meaningful entities can be reliably extracted. In this context, though OCR accuracy functions as a necessary baseline, the accuracy of entity extraction offers a more significant measure of success.

Reframing accuracy in this way also foregrounds the methodological problem of absence. Historical directories are marked not only by errors of transcription but also by omissions that no amount of post-correction can recover, pointing to gaps in the record that are not simply unknown but are unknowable under traditional evidentiary models. In doing so, I draw from recent work in archival theory and data epistemology and propose a framework for analyzing the structural conditions under which certain populations, publications, or practices fail to be preserved. Rather than treating archival silence as a dead end, I argue that the data we gain from these newspaper directories form a “speculative bibliography” that enhances the official bibliographic record and helps enable new forms of historical inference about marginal or ephemeral print cultures.¹⁶

To this end, I propose a speculative bibliography that draws from the redundancy, structure, and scale of yearly-produced directories to surface traces of the excluded. This is not recovery in the sense of restoring what has been lost, but rather remediation in the sense of reframing directories as dynamic objects that suggest what is missing.

Speculative bibliography thus offers a framework for drawing attention to silences, contextualizing fragments, and expanding bibliographic records beyond their official limits.

Speculative bibliography is less about producing a complete list and is more concerned with highlighting the contours of incompleteness. It reframes directories as evidence not only of what was recorded but also of what was deliberately or structurally excluded. By integrating these insights into scholarly editions, editors can signal both presence and absence, making visible the historical conditions under which print cultures flourished, vanished, or were obscured.

The broader contribution of this work, however, lies in the development of a reproducible, open-source, well-structured digital resource for humanists, especially those studying and stewarding large physical and digital newspaper repositories. By fine-tuning multimodal models on tabular sources, I demonstrate an approach that not only enhances OCR pipelines but also enables humanists and cultural heritage practitioners to generate machine-readable bibliographic datasets from other sources with similar formats. In doing so, it reframes digitized directories as dynamic remediated objects that can be queried, compared, and recontextualized, extending the scholarly infrastructure available for the study of newspapers and other forms of print culture.

As a concrete demonstration, the open dataset of African American titles extracted from Rowell and Ayer exemplifies how such remediation can amplify overlooked archives, offering librarians and scholars both a resource for recovery and a model for extending bibliographic records beyond their originating commercial context.

Applications for Editors

While this article is primarily methodological, its implications extend directly to the work of those concerned with scholarly editing. Directories, such as those produced by Rowell and Ayer, were themselves bibliographic instruments as they aggregated and standardized information about serials for commercial purposes. By treating these directories as tabular data and extracting structured entities, editors gain a reproducible workflow for generating metadata that supports bibliographic description, textual annotation, and contextualization.

For editors working with fragmentary or underdocumented serials, particularly those associated with marginalized communities or ephemeral print cultures, these methods offer a way to supplement scarce metadata with computationally derived structures by extending the collation of textual variants to the remediation of bibliographic records themselves and enabling editors to bring otherwise invisible or inaccessible serials into view. This alignment between data extraction and editorial practice suggests that methods developed here can enrich critical editions, digital collections, and thematic journal issues alike. By enhancing discoverability and contextualization, editors are better positioned to serve scholarly and public audiences.

Consider, for example, an editor preparing a digital edition of a nineteenth-century African American newspaper that survives in only a handful of scattered issues. The directory data extracted here could supply publication dates, frequency changes, editorial successions, and circulation figures that no surviving issue contains on its own. Cross-referenced across multiple years of Rowell and Ayer, these entities could populate editorial annotations, establish a publication timeline, and identify contemporaneous titles published in the same city. This is all contextual work that would otherwise require months of manual archival research, if it were possible at all.

Historians and archivists have long relied on these directories as a source of bibliographic information. The N. W. Ayer *Annuals*, in particular, have been used as a source of information for ethnic and non-English-language newspapers. As an advertising agent, Ayer began assigning African American designations to newspapers in 1881, and by 1910 the Ayer directory included a dedicated section listing African American newspapers as a particular subgenre.¹⁷

The history of African American newspaper bibliography illustrates both the urgency of this work, and precedents for methodological innovation. Long before computational methods were available, scholars undertook exactly the kind of speculative documentation this article proposes by assembling records of titles they could not physically locate in the hope that future researchers might find them. Historians and bibliographers have long relied on Ayer’s attention to ethnicity and language. While compiling “A Register and History of Negro Newspapers in the United States: 1827–1950,” Armistead Scott Pride noted that “no single source has been more rich in providing titles” than the *Ayer Annuals*.¹⁸ Pride wrote his “Register and History” after having scoured the country as part of an American Council of Learned Societies (ACLS)–funded project to microfilm African American newspapers that were printed prior to 1900. That microfilming project found only 467 titles and resulted in *Negro Newspapers on Microfilm: A Selected List*. This microfilm collection later became the basis for the African American titles in the Library of Congress’s Chronicling America Historic Newspapers digital collection.¹⁹

Pride quickly realized the need to document the existence of other African American newspapers, even if he could not physically locate them. He hoped that this process would help them to be located later. That documentation process resulted in his thesis, “Register and History,” in which he identified 2,700 titles.²⁰ By extracting entities from Rowell and Ayer, I have been able to build from Pride’s work to identify and extract 6,301 entries across 2,820 unique African American newspaper titles across the directories dataset. This dataset is released with this article to assist historians and bibliographers of the African American press. The gap between Pride’s 2,700 titles assembled through years of manual archival labor and the 2,820 recovered computationally underscores both the scale of what the directories incidentally preserved and the potential of entity extraction to surface records that remain invisible to traditional bibliographic methods. These 2,820 titles can be broken down into 1,754 from Ayer, 737 from Rowell, and 329 existing in both. There are a total of 9,671 records across the time period.

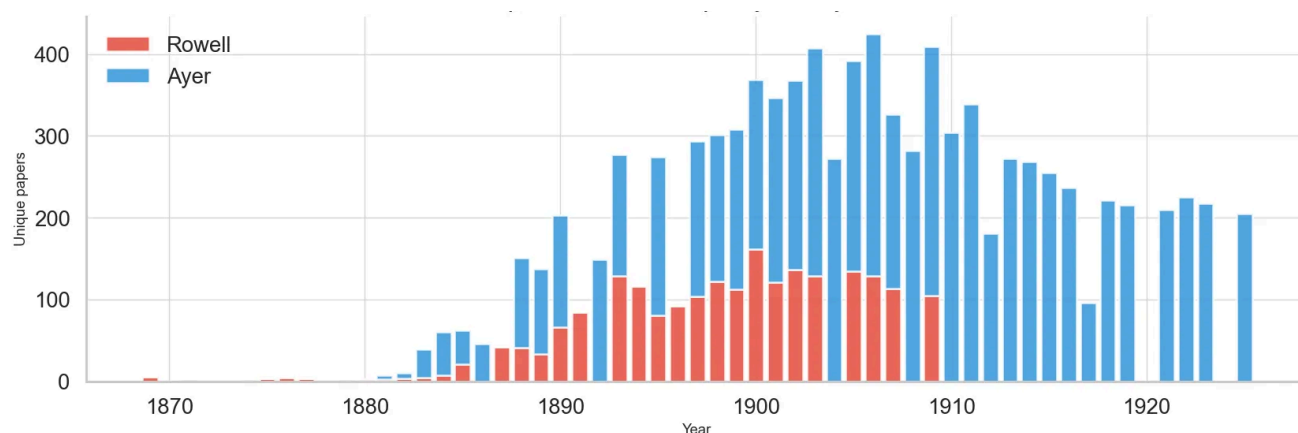


Figure 4. African American press listings by directory and year.

Although this figure may appear only modestly larger than Pride’s 2,700, the comparison is deceptive. Pride’s “Register” spanned more than a century, from 1827 to 1950, while the directory data extracted here cover a narrower window: Rowell from 1869 to 1909 and Ayer from 1880 to 1925. Additionally, Pride found only a “few titles” in Rowell’s *Directory*, whereas this work identifies 737.²¹ That a shorter time period yields a comparable number of unique titles suggests both the density of the directory records and the extent of what remains undocumented outside them.

These records exist as a “speculative bibliography,” as they are computationally created and their metadata is meant to enhance the existing record. The hope is that this dataset will draw attention to existing archival and bibliographic gaps and allow future editors to locate, contextualize, and incorporate these titles into scholarly editions and digital collections.

Conclusion

This article proposes both a technical workflow and a methodological reorientation. By treating nineteenth- and twentieth-century business directories as tabular data rather than static reference works, I show how transformer-based computer vision models can extract structured entities at a scale that manual transcription would struggle to achieve. But the more consequential claim is about what that extraction makes possible by producing a remediated dataset that transforms commercial records into scholarly infrastructure that can be queried, linked, and interrogated.

The subdataset of 2,820 African American newspaper titles drawn from Rowell’s and Ayer’s directories illustrates this potential most clearly. Where Pride’s landmark “Register” documented 2,700 titles and the subsequent ACLS microfilming project located just 467, computationally extracted directory data substantially expands the bibliographic horizon by making the shape of incompleteness more visible. In producing this speculative bibliography, I am creating a structured resource that enhances the existing record while drawing attention to its gaps.

The approach outlined here is not limited to newspaper directories. Any tabular historical source, such as city directories, trade catalogs, or institutional registers, could follow a similar workflow to yield structured datasets from documents that were never designed to be read computationally. The tools and methods are reproducible and open-source, and the hope is that other scholars, editors, and cultural heritage practitioners will adapt them to their own projects.

The dataset can be found at <https://doi.org/10.5281/zenodo.19422627>.

Acknowledgments

This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

-
1. Jessica Marie Johnson, "Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads," *Social Text* 36, no. 4 (2018): 57–79, <https://doi.org/10.1215/01642472-7145658>; Jordan E. Taylor, "Enquire of the Printer: Newspaper Advertising and the Moral Economy of the North American Slave Trade, 1704–1807," *Early American Studies* 18, no. 3 (2020): 287–323. ↵
 2. George Presbury Rowell, ed., *American Newspaper Directory* (New York: Geo. P. Rowell & Co., 1869–1909); and N. W. Ayer & Son, *American Newspaper Annual and Directory* (Philadelphia: N. W. Ayer & Son, 1910–1925). 3. For contextualization of the larger trend, see David Vincent, "The Invention of Counting: The Statistical Measurement of Literacy in Nineteenth-Century England," *Comparative Education* 50, no. 3 (2014): 266–81; and Craig Robertson, *The Filing Cabinet: A Vertical History of Information* (University of Minnesota Press, 2021). ↵
 3. For a deeper discussion of humanities data reuse, see Gregory Crane, Alison Babeu, and James Tauber, "Humanities Data Reuse: Humanity First," *Harvard Data Science Review* 7, no. 2 (2025), <https://doi.org/10.1162/99608f92.598a17b4>; Christine L. Borgman, Andrea Scharnhorst, and Milena S. Golshan, "Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse," *Journal of the Association for Information Science and Technology* 70, no. 8 (2019): 888–904, <https://doi.org/10.1002/asi.24172>; Irene V. Pasquetto, Christine L. Borgman, and Morgan F. Wofford, "Uses and Reuses of Scientific Data: The Data Creators' Advantage," *Harvard Data Science Review* 1.2 (2019), <https://doi.org/10.1162/99608f92.fc14bf2d>; and Benjamin C. G. Lee, "Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset," *Digital Humanities Quarterly* 15, no. 4 (2021), <http://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html>. ↵
 4. For a discussion of "speculative bibliography," see Ryan Cordell, "Speculative Bibliography," *Anglia* 138, no. 3 (2020): 519–31, <https://doi.org/10.1515/ang-2020-0041>. ↵
 5. Thomas G. Padilla, "Collections as Data: Implications for Enclosure," *College & Research Libraries News* 79, no. 6 (June 5, 2018), <https://doi.org/10.5860/crln.79.6.296>. ↵
 6. Andrew Piper, Chad Wellmon, and Mohamed Cheriet, "The Page Image: Towards a Visual History of Digital Documents," *Book History* 23, no. 1 (2020): 365–97, <https://doi.org/10.1353/bh.2020.0010>. ↵
 7. Daniel van Strien, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza, "Assessing the Impact of OCR Quality on Downstream NLP Tasks," INSTICC, 2020, 110, <https://www.repository.cam.ac.uk/handle/1810/304987>. ↵
 8. Van Strien et al., "Assessing the Impact," 110. ↵
 9. Piper et al., "The Page Image," 365. ↵
 10. Ian Milligan, *The Transformation of Historical Research in the Digital Age*, Elements in Historical Theory and Practice (Cambridge University Press, 2022), 18–19, <https://doi.org/10.1017/9781009026055>. ↵

11. J. P. Naiman, Morgan G. Cosillo, Peter K. G. Williams, and Alyssa Goodman, “Large Synthetic Data from the arXiv for OCR Post Correction of Historic Scientific Articles,” in *Linking Theory and Practice of Digital Libraries*, Twenty-Seventh International Conference on Theory and Practice of Digital Libraries (Springer, 2023), 265–74, https://doi.org/10.1007/978-3-031-43849-3_23; see also Van Strien et al., “Assessing the Impact”; Melissa Dell, Jacob Carlson, Tom Bryan, et al., “American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers,” *Advances in Neural Information Processing Systems* 36 (2023): 80744–80772, <http://arxiv.org/abs/2308.12477>; Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C. Dubnick, Ted Underwood, and J. Stephen Downie, “Impact of OCR Quality on BERT Embeddings in the Domain Classification of Book Excerpts,” *CEUR Workshop Proceedings*, vol. 2989 (2021), 266–79, https://ceur-ws.org/Vol-2989/long_paper43.pdf. ↵
12. Eun Seo Jo and Timnit Gebru, “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020), 306–16, <https://doi.org/10.1145/3351095.3372829>. ↵
13. Recent work in this space includes Henk Alkemade, Steven Claeysens, Giovanni Colavizza, et al., “Datasheets for Digital Cultural Heritage Datasets,” *Journal of Open Humanities Data* 9, no. 1 (2023): 17, <https://doi.org/10.5334/johd.124>; Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al., “Datasheets for Datasets,” *Communications of the ACM* 65, no. 12 (2021): 86–92, <https://doi.org/10.48550/arXiv.1803.09010>; Katherine Bode, “Why You Can’t Model Away Bias,” *Modern Language Quarterly* 81, no. 1 (2020): 95–124, <https://doi.org/10.1215/00267929-7933102>; Benjamin Charles Germain Lee, “The ‘Collections as ML Data’ Checklist for Machine Learning and Cultural Heritage,” *Journal of the Association for Information Science and Technology* 76, no. 2 (2025): 375–96, <https://doi.org/10.48550/arXiv.2207.02960>; Peter Organisciak, Boris Capitanu, Ted Underwood, and J. Stephen Downie, “Access to Billions of Pages for Large-Scale Text Analysis,” in *iConference 2017 Proceedings 2* (2017): 66–76, <https://www.ideals.illinois.edu/items/104395>. ↵
14. Molly Hardy and Lindsay DiCuirici, “Critical Cataloging and the Serials Archive: The Digital Making of ‘Mill Girls in Nineteenth-Century Print,’” *Archive Journal*, November 2019, <http://www.archivejournal.net/?p=8073>. ↵
15. Johnson, “Markup Bodies”; Taylor, “Enquire of the Printer.” ↵
16. Cordell, “Speculative Bibliography.” ↵
17. Armistead Scott Pride mentions relying heavily on Ayer’s *Annual* when compiling his thesis, “A Register and History of Negro Newspapers in the United States: 1827–1950” (Northwestern University, 1950), 9. Pride notes that Ayer was particularly valuable because he recorded African American newspapers separately beginning in 1910, something that other compilers failed to do. ↵
18. Pride also noted in “Register and History” that Ayer was not comprehensive; there were many titles, especially those from Kansas, that were not found in Ayer (9). ↵

19. For further information, see Armistead Scott Pride, *Negro Newspapers on Microfilm: A Selected List* (Washington, DC: Library of Congress, Photoduplication Service, 1953),
<https://catalog.hathitrust.org/Record/001174706>. ↵
20. Pride, “Register and History,” 3. ↵
21. Pride noted in “Register and History” that he gained “few titles” from Rowell and had difficulty in identifying any papers from 1820 to 1869 (3). Rowell began publication of his *Directory* in 1869. For context, Irvine Garland Penn did not appear to rely on any directory and documented ten African American newspapers active in 1870 and thirty-one in 1880. See I. Garland Penn, *The Afro-American Press and Its Editors* (Springfield, MA: Willey, 1891), 112–14. ↵