LIBRARY OF CONGRESS

April 1, 2026

Chronicling America as a Large Dataset for Newspaper Research

University of Georgia, University Libraries

Robin Pike
NDNP Coordinator,
Head, Digital Collections Services Section
Serial & Government Publications Division

Poll

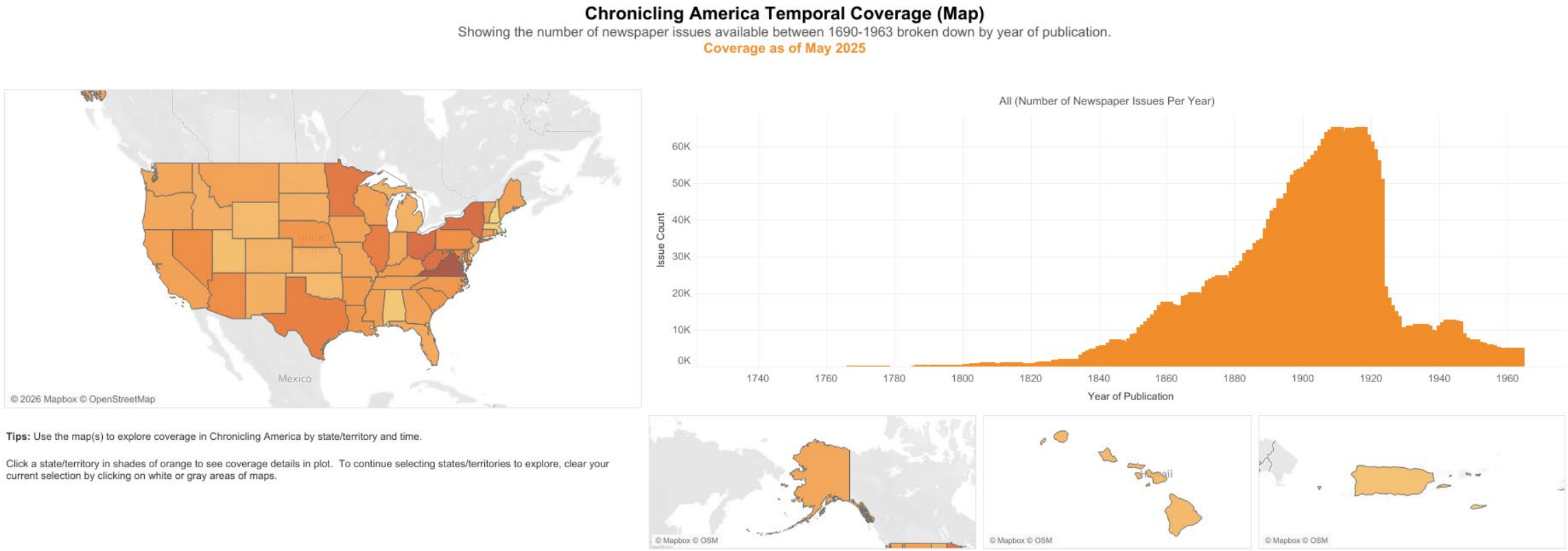
- Who has used historical newspapers?
- Who has used Chronicling America?
- Who has used Chronicling America as data?

ABOUT THE DATA

What is Chronicling America?

- Chronicling America
- Free, publicly accessible database of newspapers
- 1736-1963
- +23.7 million pages
- +4,500 newspaper titles
- 50 states, DC, PR, VI
- 34 ethnicities
- 34 languages

Chronicling America: Coverage



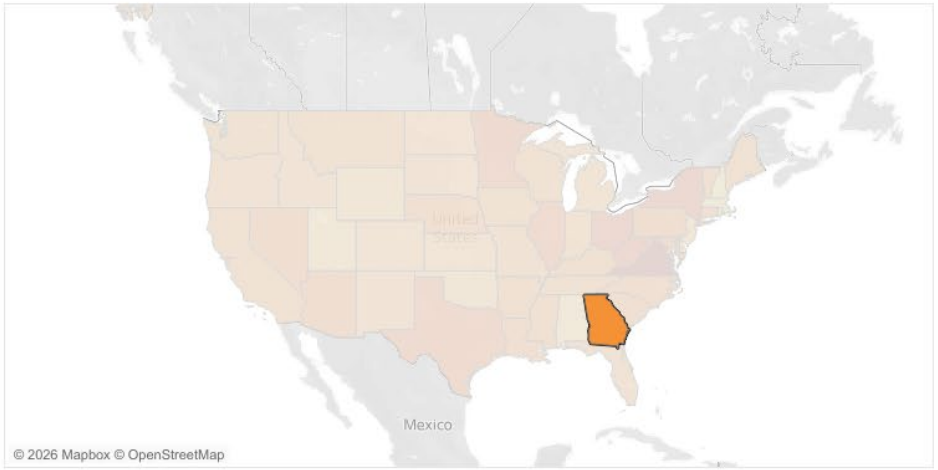
http://www.loc.gov/ndnp/data-visualizations/

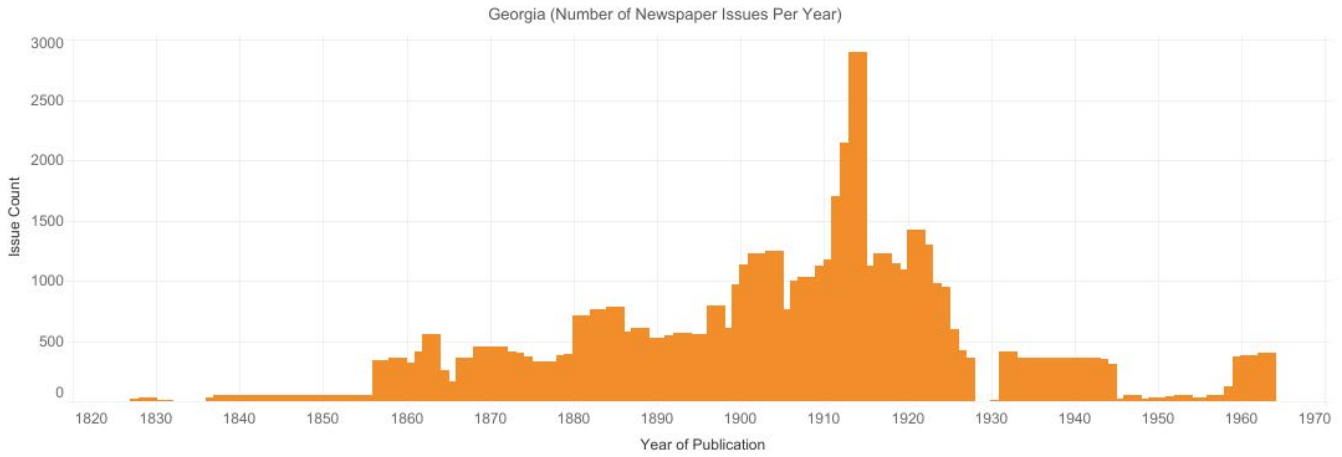
Georgia: Coverage

Chronicling America Temporal Coverage (Map)

Showing the number of newspaper issues available between 1690-1963 broken down by year of publication.

Coverage as of May 2025





Tips: Use the map(s) to explore coverage in Chronicling America by state/territory and time.

Click a state/territory in shades of orange to see coverage details in plot. To continue selecting states/territories to explore, clear your current selection by clicking on white or gray areas of maps.

About: Data used in this visualization is from https://www.loc.gov/collections/chronicling-america/ as of 5/1/2025.
More info at: https://www.loc.gov/ndnp/data-visualizations/.

Maps & Visualizations: https://www.loc.gov/ndnp/data-visualizations/

Understanding Historical Newspapers

- Optical Character Recognition (OCR)
- Stories evolve over time
- Historical biases, knowledge

- Insert graphic of something

Evolution of Language

Historical Vocabulary

- Suffrage vs. Women's rights
- Consumption vs. White Plague vs. Tuberculosis
- Negro vs. Colored vs. African American vs. Black

Historical Spellings

- Place names
- Misspellings, multiple spellings common
- Aeroplane vs. Airplane

EXPLORE THE DATA

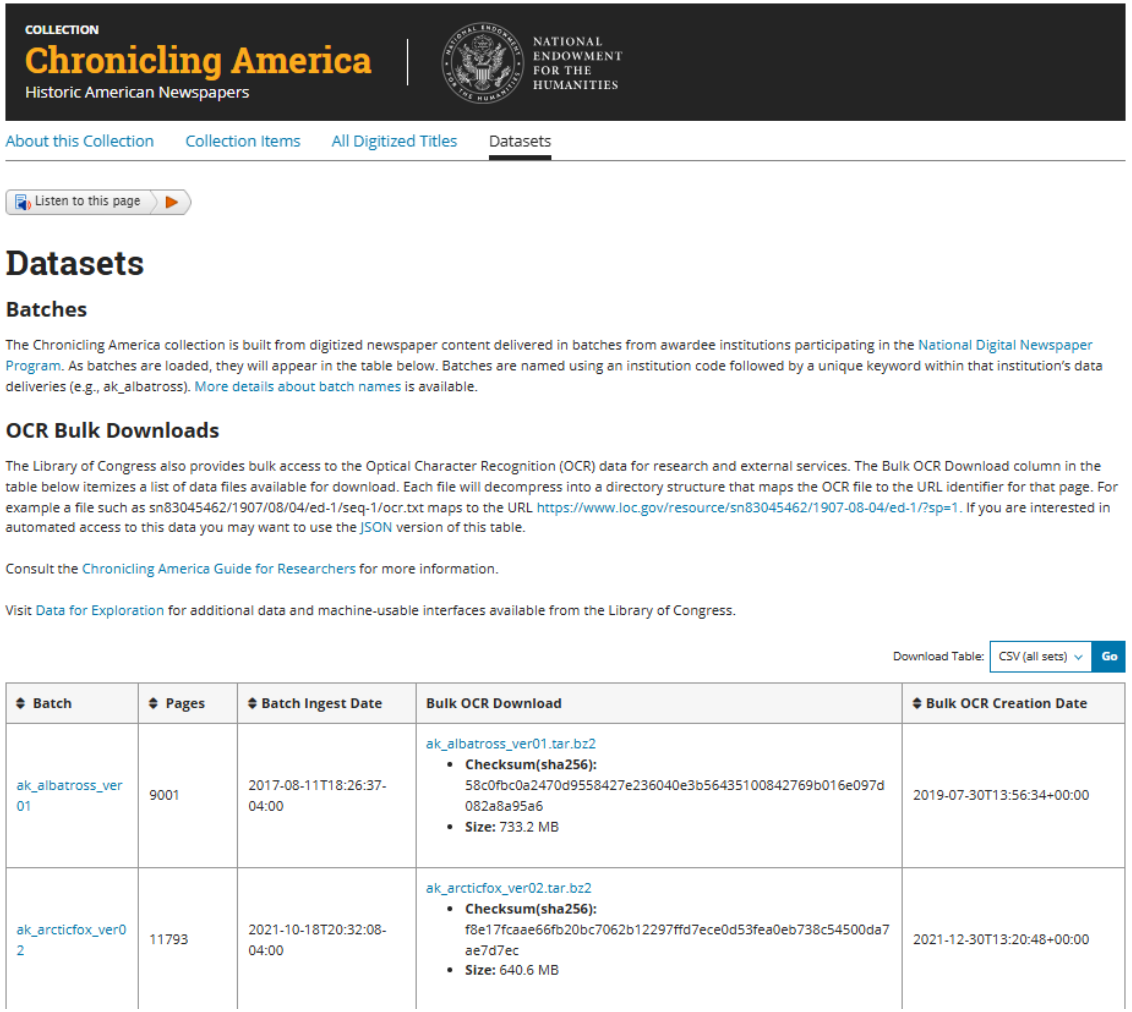
Packages of Chronicling America Data

- NEW! Datasets Tab
  - Batch Data
    - Images, OCR, metadata
  - Bulk OCR
    - OCR, minimal metadata
- API search, download queries
- Directory of U.S. Newspapers in American Libraries
  - Catalog record, digitized and undigitized newspapers

- Selected Data
  - Newspaper title data
    - Catalog records, history essays
  - Coming Later: Image datasets
    - From Newspaper Navigator

- Download data vs. data subset vs. API?

Batch and OCR: https://www.loc.gov/collections/chronicling-america/datasets/



COLLECTION
Chronicling America
Historic American Newspapers

NATIONAL ENDOWMENT FOR THE HUMANITIES

About this Collection | Collection Items | All Digitized Titles | Datasets

Listen to this page

Datasets

Batches

The Chronicling America collection is built from digitized newspaper content delivered in batches from awardee institutions participating in the National Digital Newspaper Program. As batches are loaded, they will appear in the table below. Batches are named using an institution code followed by a unique keyword within that institution's data deliveries (e.g., ak_albatross). More details about batch names is available.

OCR Bulk Downloads

The Library of Congress also provides bulk access to the Optical Character Recognition (OCR) data for research and external services. The Bulk OCR Download column in the table below itemizes a list of data files available for download. Each file will decompress into a directory structure that maps the OCR file to the URL identifier for that page. For example a file such as sn83045462/1907/08/04/ed-1/seq-1/ocr.txt maps to the URL https://www.loc.gov/resource/sn83045462/1907-08-04/ed-1/?sp=1. If you are interested in automated access to this data you may want to use the JSON version of this table.

Consult the Chronicling America Guide for Researchers for more information.

Visit Data for Exploration for additional data and machine-usable interfaces available from the Library of Congress.

Download Table: CSV (all sets) Go

Batch	Pages	Batch Ingest Date	Bulk OCR Download	Bulk OCR Creation Date
ak_albatross_ver01	9001	2017-08-11T18:26:37-04:00	ak_albatross_ver01.tar.bz2 • **Checksum(sha256):** 58c0fbc0a2470d9558427e236040e3b56435100842769b016e097d082a8a95a6 • **Size:** 733.2 MB	2019-07-30T13:56:34+00:00
ak_arcticfox_ver02	11793	2021-10-18T20:32:08-04:00	ak_arcticfox_ver02.tar.bz2 • **Checksum(sha256):** f8e17fcaae66fb20bc7062b12297ffd7ece0d53fea0eb738c54500da7ae7d7ec • **Size:** 640.6 MB	2021-12-30T13:20:48+00:00

About LOC.GOV API

- Library of Congress Application Programming Interface (API)
  - Download collection content files and structured data (JSON/YAML) about collections
- 2025 Chronicling America collection accessible via loc.gov API
- Benefits:
  - Works across collections
  - Cleaner, modern data structures
  - Long-term stability
- Publicly accessible: no API key and authentication required (for now)
- Rate limit: 20 requests/minute or 3 seconds/request (JSON/YAML)
- Documentation

Jupyter Notebooks

- Chronicling America Notebooks
  - ReadMe
  - Six Jupyter Notebook Examples
    - Created based on Common Research Questions received through ask.loc.gov
- Combine step-by-step explanations with executable Python code
- Run in environments like Anaconda or Google Collab
- Use Cases:
  - Digital Humanities Research
  - Text Mining
  - Data Visualization

Example: Word Frequency, Geographic Use

Using Chronicling America to analyze word frequency and geographic usage

Feel free to download this notebook and put in your own search queries.

Notebook Example

For this example, we will look at the term "**influenza**" and its occurence in the U.S. newspapers in Chronicling America during **1800** and **1830**.

Specifically, we want to utilize the API to look at the word usage in relationship to time and location:

1. Time: When was the term published in newspapers?
2. Location: Where was the term most commonly used based on a newspaper's publication location?

Importing Modules [Required]

The following imports are required for the scripts to run properly:

---

1. Run the following code below.
   - It will import all the modules you need for this notebook.
   - Do not change anything.

In [1]:

```
import time
import re
import json
from urllib.request import urlopen
import requests
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import pprint
```

Title Record

NEWSPAPER

The Athens Republique (Athens, Ga.) 1919-????

About this Newspaper | Libraries that Have It | Browse Digitized Issues

About The Athens Republique (Athens, Ga.) 1919-????

Recent World War I veteran Julian Lucasse Brown began publishing the *Athens Republique* in November 1919 in Athens, Georgia. According to its masthead, the paper was "Devoted to the Religious, the Economic, and the Industrial Development of the Colored Race." The *Republique* was also the official organ for the local Jeruel Baptist Association, which ran the Jeruel Academy, a private school for Black students in the city. The *Republique* circulated weekly on Saturdays and covered stories on the affairs of the African American community in Athens. The paper also regularly featured national reports of efforts to fight the Ku Klux Klan and lynchings across the country. Additionally, Brown devoted multiple pages of each issue to societal news in Athens and surrounding towns, including weddings, deaths, illnesses, and church events often ignored by the white-run press.

In 1923, Brown relocated the paper to an office on Hull Street in an area known as the "Hot Corner" in downtown Athens. The Hot Corner supported a thriving African American business community and was home to the celebrated Morton Theatre, one of the first black-owned and operated vaudeville theatres in the country. Brown supplemented his journalistic endeavor by working as a notary, serving as secretary of the Allied National Farm Association (also headquartered on Hull Street), and selling printed materials out of his office. By 1927, the *Athens Republique* was no longer in business. In the decade that followed, Brown and his wife, Katherine, moved to Alabama, where he served as a teacher and printer at the Tuskegee Institute.

Show More

About this Newspaper

Title
The Athens Republique (Athens, Ga.) 1919-????

Names
Brown, Julian L., editor
Jeruel Baptist Association

Dates of Publication
1919-????

Created / Published
Athens, Ga. : Athens Republique

Headings
- African Americans--Georgia--Athens--Newspapers
- Noirs--Géorgie (État)--Athens--Journaux
- Noirs américains--Géorgie (État)--Athens--Journaux
- African Americans
- Georgia--Athens

Availability
View All Front Pages »
Check the "Libraries that Have It" tab for additional newspaper issues, or, if present, select the LCCN Permalink for more LC holdings

Part of
The Athens Republique (Athens, Ga.) 1919-???? (20)
Directory of U.S. Newspapers in American Libraries (158,740)
Library of Congress Online Catalog (1,759,536)
Chronicling America (3,198,586)
Serial and Government Publications Division (3,219,635)

Title Essay

NEWSPAPER

The Athens Republique (Athens, Ga.) 1919-????

About this Newspaper | Libraries that Have It | Browse Digitized Issues

About The Athens Republique (Athens, Ga.) 1919-????

Recent World War I veteran Julian Lucasse Brown began publishing the *Athens Republique* in November 1919 in Athens, Georgia. According to its masthead, the paper was "Devoted to the Religious, the Economic, and the Industrial Development of the Colored Race." The *Republique* was also the official organ for the local Jeruel Baptist Association, which ran the Jeruel Academy, a private school for Black students in the city. The *Republique* circulated weekly on Saturdays and covered stories on the affairs of the African American community in Athens. The paper also regularly featured national reports of efforts to fight the Ku Klux Klan and lynchings across the country. Additionally, Brown devoted multiple pages of each issue to societal news in Athens and surrounding towns, including weddings, deaths, illnesses, and church events often ignored by the white-run press.

In 1923, Brown relocated the paper to an office on Hull Street in an area known as the "Hot Corner" in downtown Athens. The Hot Corner supported a thriving African American business community and was home to the celebrated Morton Theatre, one of the first black-owned and operated vaudeville theatres in the country. Brown supplemented his journalistic endeavor by working as a notary, serving as secretary of the Allied National Farm Association (also headquartered on Hull Street), and selling printed materials out of his office. By 1927, the *Athens Republique* was no longer in business. In the decade that followed, Brown and his wife, Katherine, moved to Alabama, where he served as a teacher and printer at the Tuskegee Institute.

Provided By: Digital Library of Georgia, a project of GALILEO located at the University of Georgia Libraries

Show Less

PROJECT EXAMPLES

Applicable Research Fields

- NDNP Extras
- America's Public Bible
- American Lynching: Uncovering a Cultural Narrative
- American Stories (LLM and text layout model)
- An Epidemiology of Information: Data Mining the 1918 Influenza Epidemic Project Report
- An Industrial West? A Mixed-Methods Analysis of Newspapers Discourses about Technology over One Hundred and Ten Years (1830-1940)
- Gender Norms Do Not Persist But Converge Across Time
- Newspaper Navigator
- Viral Texts

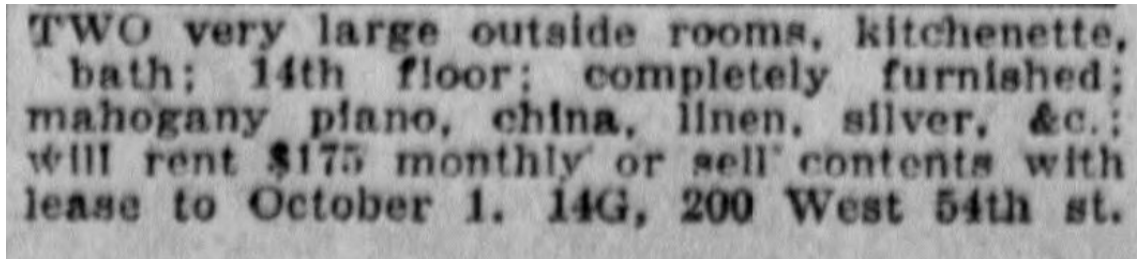
Research Questions

- “I am researching the construction of masculinities and vulnerability in the nineteenth century. I have found a list of newspapers in which the terms "manly health" appear... Is there any way of saving all the images/results and downloading them?”

- “I use comparative historical and computational methods to study the evolution of collective identity in the United States over time. I am starting a project using Help Wanted ads in Chronicling America. How do I perform a bulk download of the results of my query?”

Historical NYC Rent Data - AI Pilot Project

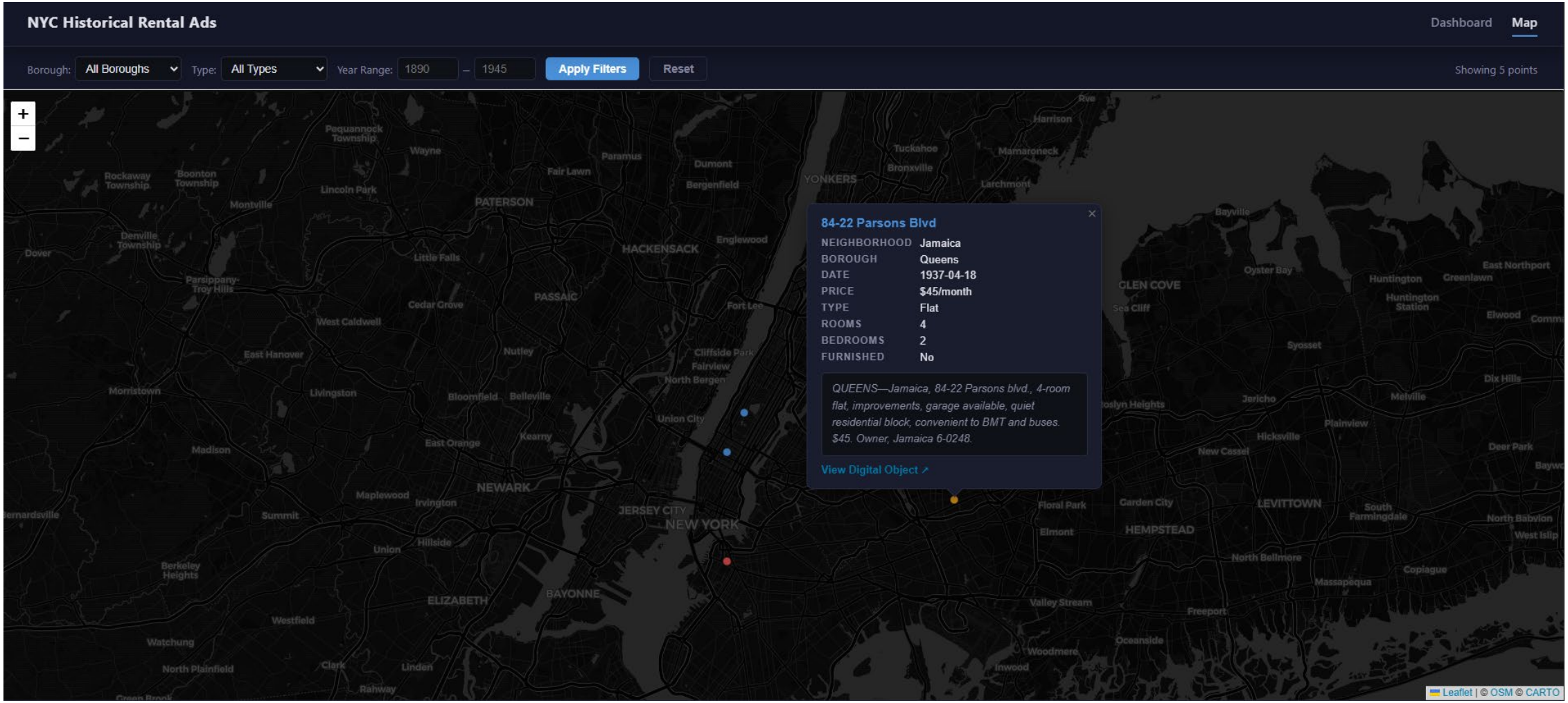
- Find apartment rental data in NYC newspapers
- OCR for 230K NYC pages with word “rent” downloaded via API
- 9,000 pages processed with AI prompt (to date)
- Prompt creates JSON output
  - Supplies latitude and longitude for address
  - AI guessing: $17.\” becomes $175 for price
- Processing will end 3/31
- Will become example Jupyter Notebook to train Library staff


TWO very large outside rooms, kitchenette, bath; 14th floor; completely furnished; mahogany piano, china, linen, silver, &c.; will rent $175 monthly or sell contents with lease to October 1. 14G, 200 West 54th st.

The New York herald (New York, N.Y.), April 30, 1922, image 84
https://www.loc.gov/resource/sn83045774/1922-04-30/ed-1/?sp=84&st=image&r=0.377,1.165,0.16,0.106,0

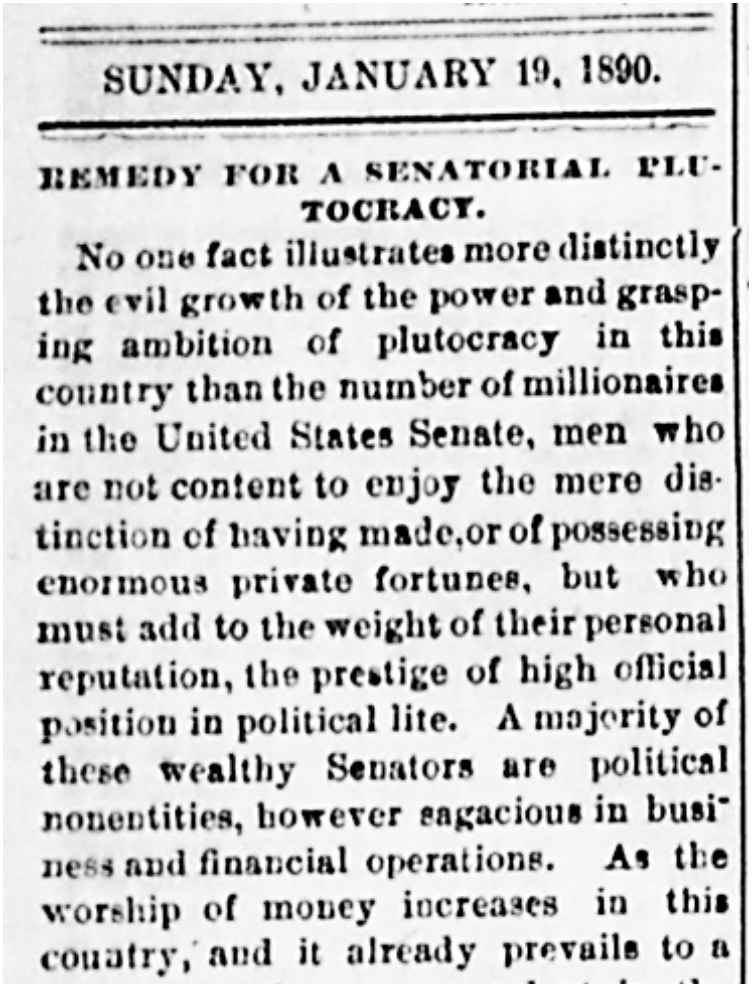

```
"address": "200 West 54th St.",
   "price": "$175",
   "building_name": null,
   "total_room_count": "2",
   "bedroom_count": null,
   "sq_ft": null,
   "is_furnished": true,
   "includes_board": false,
   "rental_period": "month",
   "type": "apartment",
   "amenities": ["kitchenette", "bath", "mahogany piano", "china",
"linen", "silver"],
   "full_text": "TWO very large outside rooms, kitchenette, bath:
14th floor; completely furnished; i mahogany piano, china, linen,
silver, \\<'tl1 rent $17.\" monthly or sell contents with leaao to
October 1. J4G, 200 West 54Ui st.",
   "neighborhood": "Midtown",
   "borough": "Manhattan",
   "latlong": "40.7640,-73.9820"
```

Historical NYC Rent Data - AI Pilot Project



IMPROVEMENTS AND RESOURCES

OCR Reprocessing



SUNDAY, JANUARY 19, 1890.

REMEDY FOR A SENATORIAL PLUTOCRACY.

No one fact illustrates more distinctly the evil growth of the power and grasping ambition of plutocracy in this country than the number of millionaires in the United States Senate, men who are not content to enjoy the mere distinction of having made,or of possessing enormous private fortunes, but who must add to the weight of their personal reputation, the prestige of high official position in political lite. A majority of these wealthy Senators are political nonentities, however sagacious in busi- ness and financial operations. As the worship of money increases in this couatry, and it already prevails to a

Image Source: The times (Richmond, Va.), January 19, 1890

Before:

t «???????? ini ?: «? t I » ' la ihn miluoaaitea il wi,«. » but WBO weight OÍ 11 «·. ? iiiii.i . ? Inni" ? « tie tini ind it air. :» to ?» ill tin lbs muí

After:

SUNDAY. JANUARY 19, 1800.
REMEDY SENATORIAL PLU
No one fact illustrates more distinctly the evil growth of the power and grasping ambition of plutocracy in this country than the number of millionaires in the United States Senate, men who are not content to enjoy the mere dis tinction of having made,or of possessing enormous private fortunes, but who must add to the weight of their personal reputation, the prestige of high official position in political lite. A majority of these wealthy Senators are political nonentities, however sagacious in busi” ness and financial operations. As the worship of movey iocreases in this couafry, and it already prevails to a

Reprocessed Batches

Online as of 4/1/2026:

- ~245,000 reprocessed pages
- 52 batches
- Many more on the way!

Improved Machine-Readable Text for Newspapers

The Library of Congress has launched a new project to improve the text supporting keyword searches in Chronicling America. Read about our recent work and future plans to make this valuable historical collection easier to explore.

Total Reprocessed Pages to Date: 189,058

Background | What is OCR? | NDNP-Open-OCR

Background

Chronicling America provides access to historic newspapers digitized under the National Digital Newspaper Program (NDNP). Sponsored by the Library of Congress and the National Endowment for the Humanities (NEH), the NDNP began in 2005. In anticipation of the NDNP's 20th year, the Library launched an effort to make the digitized newspaper data more accessible to users by reprocessing select newspaper content digitized prior to 2012 to improve its machine-readable text.

Machine-readable text is created by a technology called Optical Character Recognition (OCR). Using the Tesseract Open Source OCR Engine and customized processing steps, the Library created a new OCR reprocessing pipeline called NDNP-Open-OCR.

More information on this project is coming soon.

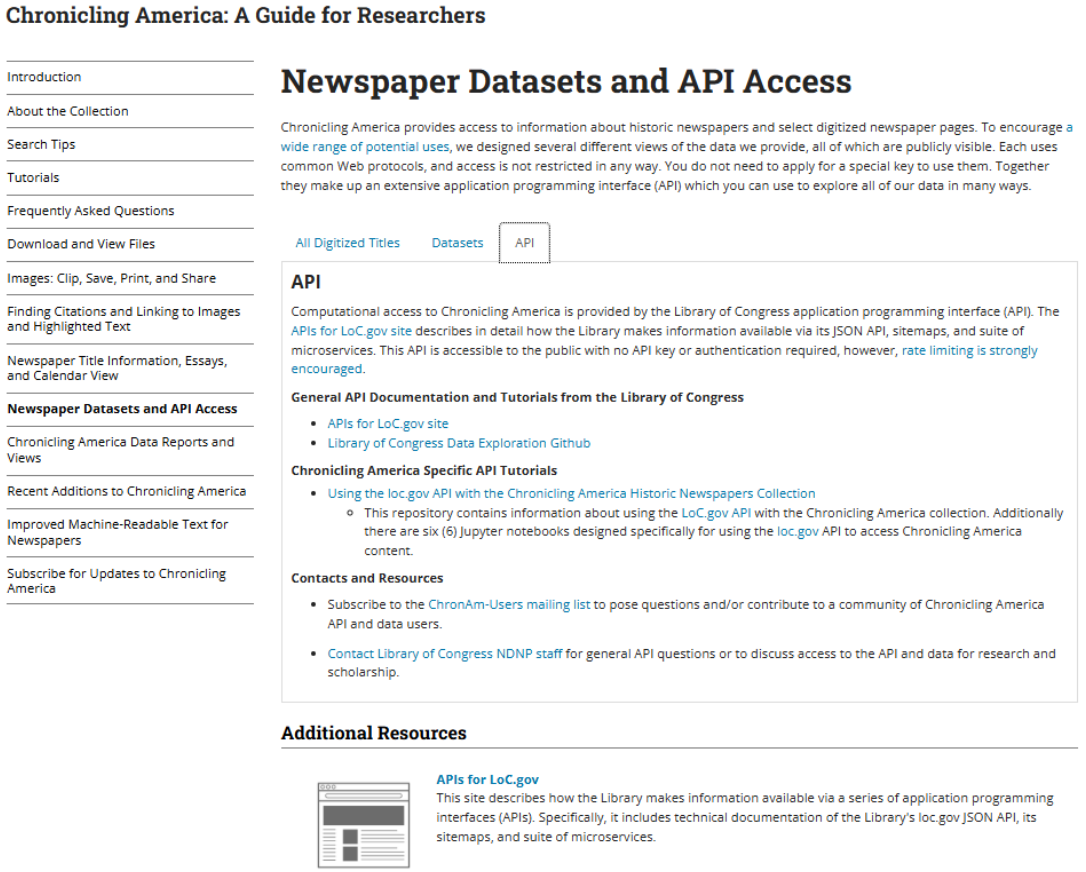
For questions, please contact ndnptech@loc.gov.

Date Reprocessed Batch Added	Contributor	Batch Name	Page Count	Content on Batch
2025-09-26	VA - Library of Virginia; Richmond, VA	vi_kiss_ver02	3313	Richmond Dispatch (sn85038614) 1895-1899
2025-09-26	VA - Library of Virginia; Richmond, VA	vi_styx_ver03	813	Highland Recorder (sn95079246) 1893-1896
2025-09-26	VA - Library of Virginia; Richmond, VA	vi_yes_ver03	1490	Richmond Planet (sn84025841) 1889-1899
2025-09-25	VA - Library of Virginia; Richmond, VA	vi_heart_ver02	6739	Richmond Dispatch (sn85038614) 1888-1895
2025-09-25	VA - Library of Virginia; Richmond, VA	vi_journey_ver02	5862	The Daily Dispatch (sn84024738) 1880-1884
2025-04-11	LC- Library of Congress, Washington, DC	dlc_avanti_ver02	2374	New-York Tribune (sn83030214) 1880
2025-04-11	LC- Library of Congress, Washington, DC	dlc_buick_ver02	2376	New-York Tribune (sn83030214) 1882

https://guides.loc.gov/chronicling-america/improved-text

Research Guides

- Chronicling America: A Guide for Researchers
  - Newspaper Datasets and API Access
- Directory of U.S. Newspapers in American Libraries Research Guide
- Datasets at the Library of Congress
- APIs for LoC.gov



Chronicling America: A Guide for Researchers

Introduction
About the Collection
Search Tips
Tutorials
Frequently Asked Questions
Download and View Files
Images: Clip, Save, Print, and Share
Finding Citations and Linking to Images and Highlighted Text
Newspaper Title Information, Essays, and Calendar View
Newspaper Datasets and API Access
Chronicling America Data Reports and Views
Recent Additions to Chronicling America
Improved Machine-Readable Text for Newspapers
Subscribe for Updates to Chronicling America

Newspaper Datasets and API Access

Chronicling America provides access to information about historic newspapers and select digitized newspaper pages. To encourage a wide range of potential uses, we designed several different views of the data we provide, all of which are publicly visible. Each uses common Web protocols, and access is not restricted in any way. You do not need to apply for a special key to use them. Together they make up an extensive application programming interface (API) which you can use to explore all of our data in many ways.

All Digitized Titles | Datasets | API

API

Computational access to Chronicling America is provided by the Library of Congress application programming interface (API). The APIs for LoC.gov site describes in detail how the Library makes information available via its JSON API, sitemaps, and suite of microservices. This API is accessible to the public with no API key or authentication required, however, rate limiting is strongly encouraged.

General API Documentation and Tutorials from the Library of Congress

- APIs for LoC.gov site
- Library of Congress Data Exploration Github

Chronicling America Specific API Tutorials

- Using the loc.gov API with the Chronicling America Historic Newspapers Collection
  - This repository contains information about using the LoC.gov API with the Chronicling America collection. Additionally there are six (6) Jupyter notebooks designed specifically for using the loc.gov API to access Chronicling America content.

Contacts and Resources

- Subscribe to the ChronAm-Users mailing list to pose questions and/or contribute to a community of Chronicling America API and data users.
- Contact Library of Congress NDNP staff for general API questions or to discuss access to the API and data for research and scholarship.

Additional Resources

APIs for LoC.gov
This site describes how the Library makes information available via a series of application programming interfaces (APIs). Specifically, it includes technical documentation of the Library's loc.gov JSON API, its sitemaps, and suite of microservices.

Staff

Robin Butterhof

Mike Saelee

Nathan Yarasavage

Thanks!

Robin Pike

NDNP Coordinator
Head, Digital Collections Services Section
rpike@loc.gov

Subscribe: chronam-users@listserv.loc.gov

ask.loc.gov
ndnptech@loc.gov