Bulking Up: How Accepted Standards and Evolving Technology

Advance Research in Chronicling America

Nathan Yarasavage, Deborah Thomas & Georgia Higley
Library of Congress

Chronicling America: Historic American Newspapers
http://chroniclingamerica.loc.gov/

LIBRARY OF CONGRESS

ASK A LIBRARIAN   DIGITAL COLLECTIONS   LIBRARY CATALOGS

Search [Search Loc.gov] **GO**

NATIONAL ENDOWMENT FOR THE **humanities**

CHRONICLING AMERICA
Historic American Newspapers

Search America's historic newspaper pages from 1836-1922 or use the U.S. Newspaper Directory to find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities⧉ and the Library of Congress. Learn more »

| Search Pages | **Advanced Search** | All Digitized Newspapers 1836-1922 |

US Newspaper Directory, 1690-Present

close X

Select State(s):
```
All states
Alabama
Arizona
California
Colorado
District of Columbia
Florida
Georgia
```

Or Select Newspaper(s):
```
All newspapers
Abbeville progress. (Abbeville, Vermilion Parish, La.)
Aberdeen herald. (Aberdeen, Chehalis County, W.T.)
The Abilene reflector. (Abilene, Kan.)
Abilene weekly reflector. (Abilene, Kan.)
The Abingdon Virginian. (Abingdon [Va.])
The Adair County news. (Columbia, Ky.)
Adams County news. (Ritzville, Wash.)
```

Select Year(s)*
Newspaper pages are available for newspapers published between **1836-1922***

⦿ from [1836 ▼] to [1922 ▼]

Or Date Range
○ from [   ] to [   ]

Limit Search: ☑ only front page or Specific page [   ]

Language [All ▼]

Enter Search

...with **any** of the words:
[   ]

...with **all** of the words:
[   ]

...with the **phrase**:
[   ]

...with the words:
[crowds salt lake] within [5 ▼] words of each other

Clear   **Search**

LIBRARY OF CONGRESS

ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS

Search | Search Loc.gov | GO

The Library of Congress > Chronicling America > The Salt Lake tribune.

NATIONAL ENDOWMENT FOR THE
Humanities

CHRONICLING AMERICA
Historic American Newspapers

Search America's historic newspaper pages from 1836-1922 or use the U.S. Newspaper Directory to find information about American newspapers published between 1690-present. Chronicling America is sponsored jointly by the National Endowment for the Humanities and the Library of Congress. Learn more »

Search Pages | **Advanced Search** | **All Digitized Newspapers 1836-1922**

US Newspaper Directory, 1690-Present

All states | + | from | 1836 | to | 1922 | + | enter one or more search words | GO

Pages Available: 6,691,983

Print Subscribe Share/Save Give Feedback

Browse Issues: The Salt Lake tribune.

Salt Lake City, Utah (1890-current)

Browse Issues | About | Libraries that Have It | MARC Record

Issues for: 1904

Show all front pages

Single edition:
dates in **bold**.

Multiple editions:
dates in ***bold italics***.

January, 1904
S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

February, 1904
S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29					

March, 1904
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

April, 1904
S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

May, 1904
S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

June, 1904
S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

July, 1904
S	M	T	W	T	F	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

August, 1904
S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

September, 1904
S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

October, 1904
S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

November, 1904
S	M	T	W	T	F	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30			

December, 1904
S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

THE San Francisco CALL

RUSSIA'S REPLY TO JAPAN REJECTED BY THE COUNCIL OF THE ELDER STATESMEN

ENLEAGUED REPUBLICS STIR FEAR

Four-Fifths of Navy May Be Sent to Panama.

Gathering of the Fleets an Affair of First Significance.

Says Naval Officer Hudson That La Blanca at the Isthmus Is Inadequate.

EMBEZZLER McKOWEN STAYS IN PRISON WHILE EXPERTS INVESTIGATE HIS METHODS

OFFICERS KILL SON AND SIRE

Duel Is Fought in a Room in Los Angeles.

Mikado Will Send an Army Into Korea.

ANARCHIST CONSPIRACY UNEARTHED

Texas Agitator Plots Against Nation's Noted Men.

CUBAN BILL IS PROLIFIC OF TROUBLE

Other Nations Seek to Profit by the Treaty.

Engages Attorney and Becomes Reticent.

AIRSHIP'S FLIGHT IS A SUCCESS

Craft Sails Skyward Without Aid of a Balloon.

THOUSANDS HAVE KIDNEY TROUBLE AND DON'T KNOW IT

To Prove that Swamp-Root, the Great Kidney Remedy, Will Do for YOU, Every Reader of "The Call" May Have a Sample Bottle Sent Free by Mail.

ARMY OF COLOMBIANS IS ENCAMPED NEAR THE PASSES LEADING TO PANAMA

SENATORS IN BITTER COLLOQUY

Hoar Calls Up His Resolution Asking for Information to This Government Part in the Panama Revolution.

CUSTOM-HOUSE BIDS OPENED

SENATOR HOAR'S LONG MEMORY

FRIAR LANDS FOR FILIPINOS

Facebook

Twitter

Delicious

Digg

Google

Diigo

MySpace

Reddit

StumbleUpon

Technorati

Yahoo!

LinkedIn

Google+

Tumblr

Privacy Policy

DISCLAIMER

UNIVERSITY *of* HAWAI'I
MĀNOA

Primo Beer Ads

UH Manoa Library

24 Photos   **4** Views

View all sets   **Photos**

Back to Browse All Available Issues | **Previous Issue** | **Next Issue** | **Show all front pages**
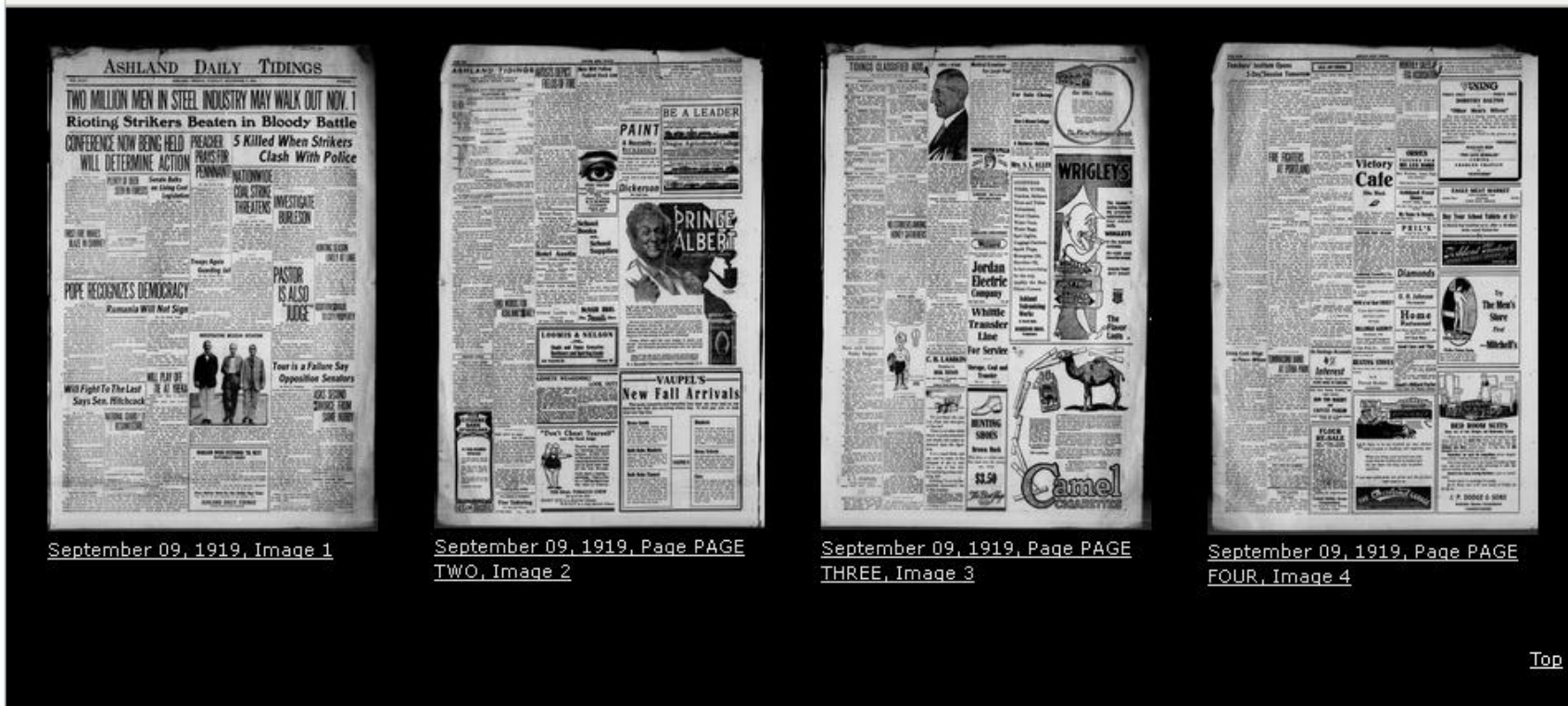
September 09, 1919, Image 1

September 09, 1919, Page PAGE TWO, Image 2

September 09, 1919, Page PAGE THREE, Image 3

September 09, 1919, Page PAGE FOUR, Image 4

Top

nys historic newspapers

Search Pages | Advanced Search | All Digitized Newspapers 1803-2010

All Cities ▾ + All counties ▾ + from 1803 ▾ to 2010 ▾ +

enter one or more search words | GO

Pages Available: 2,368,630

Contribute

Contribute To This Project

Subscribe

@

Subscribe to e-mail updates

More resources

› New York Heritage

› New York Heritage Digitized Newspapers

› New York State Newspaper Holdings (Microfilm)

NY3Rs

100 Years Ago This Week: January 24, 1914 (issues)
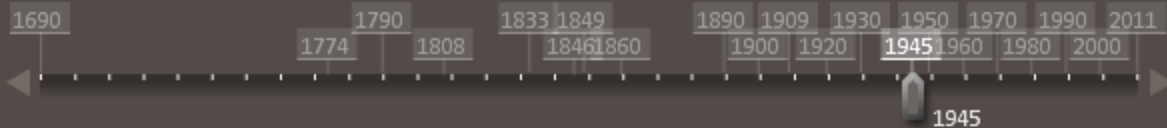
The Adirondack news. (4pp.)
St. Regis Falls, N.Y.

Watertown re-union. (8pp.)
Watertown, N.Y.

Plattsburgh daily press. (6pp.)
Plattsburgh, N.Y.

The Growth of Newspapers Across the U.S.: 1690-2011

1690   1790   1833   1849   1890   1909   1930   1950   1970   1990   2011
1774   1808   1846   1860   1900   1920   **1945**   1960   1980   2000

1945

ZOOM: 1.1x

1945
World War II

14083 listings

World War II sapped rural newsrooms of manpower as young men enlisted in droves after the Pearl Harbor attacks in 1941. As in many other industries, women filled the gap in large numbers, and legal protections were passed for newspapers that needed to suspend publication during the War. Newspapers placed new emphasis on maps and information graphics to help explain the complex and far flung battlefields to the public. Back home, the War Relocation Authority permitted Japanese-Americans to publish their own newspapers in detention camps like Manzanar in California and Gila River in Arizona.

« Back: 1930 | Next: 1950 »
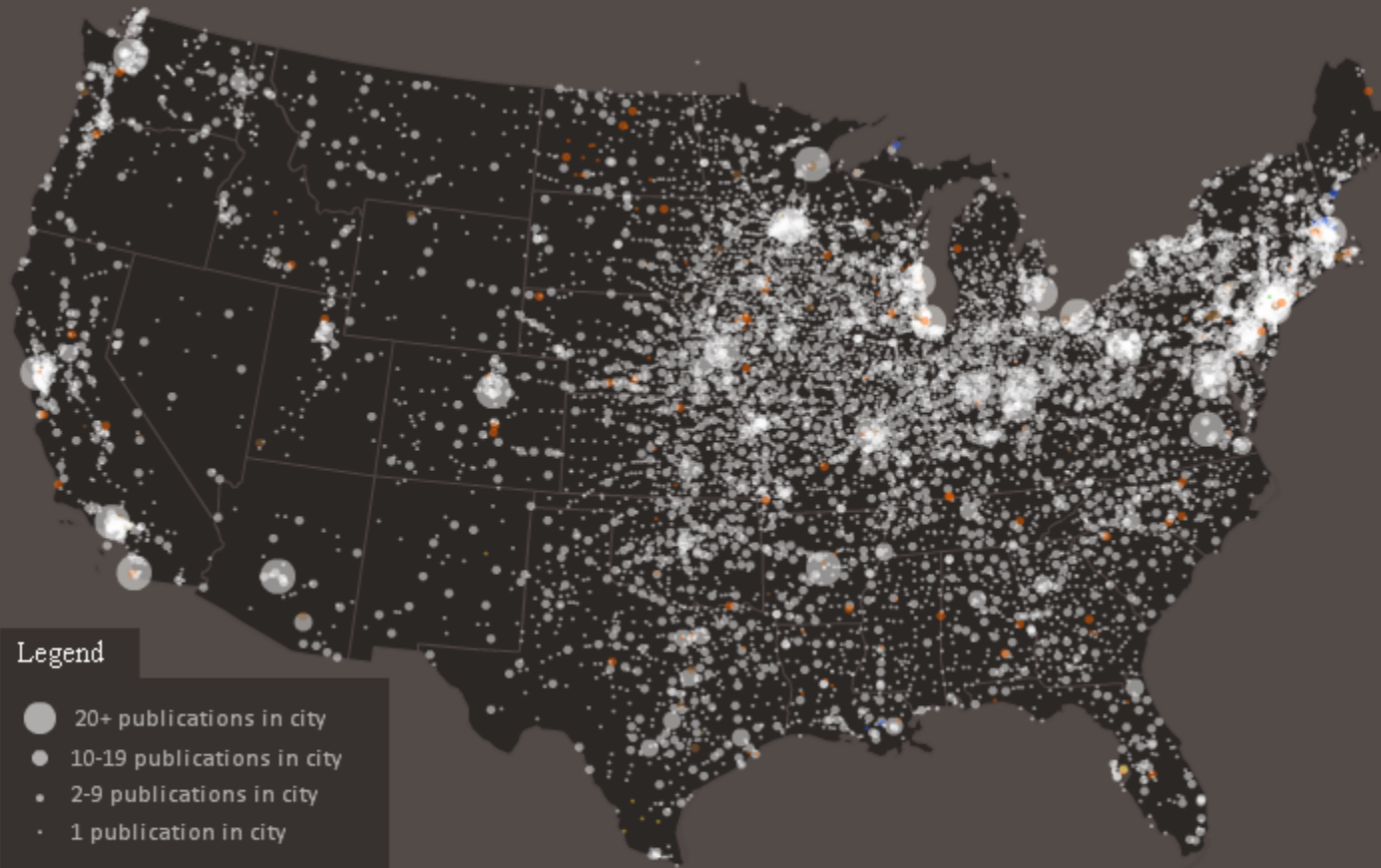
Legend

- ●   20+ publications in city
- ●   10-19 publications in city
- •   2-9 publications in city
- ·   1 publication in city

▼ Hide details (click listings to open newspaper info in a new window)

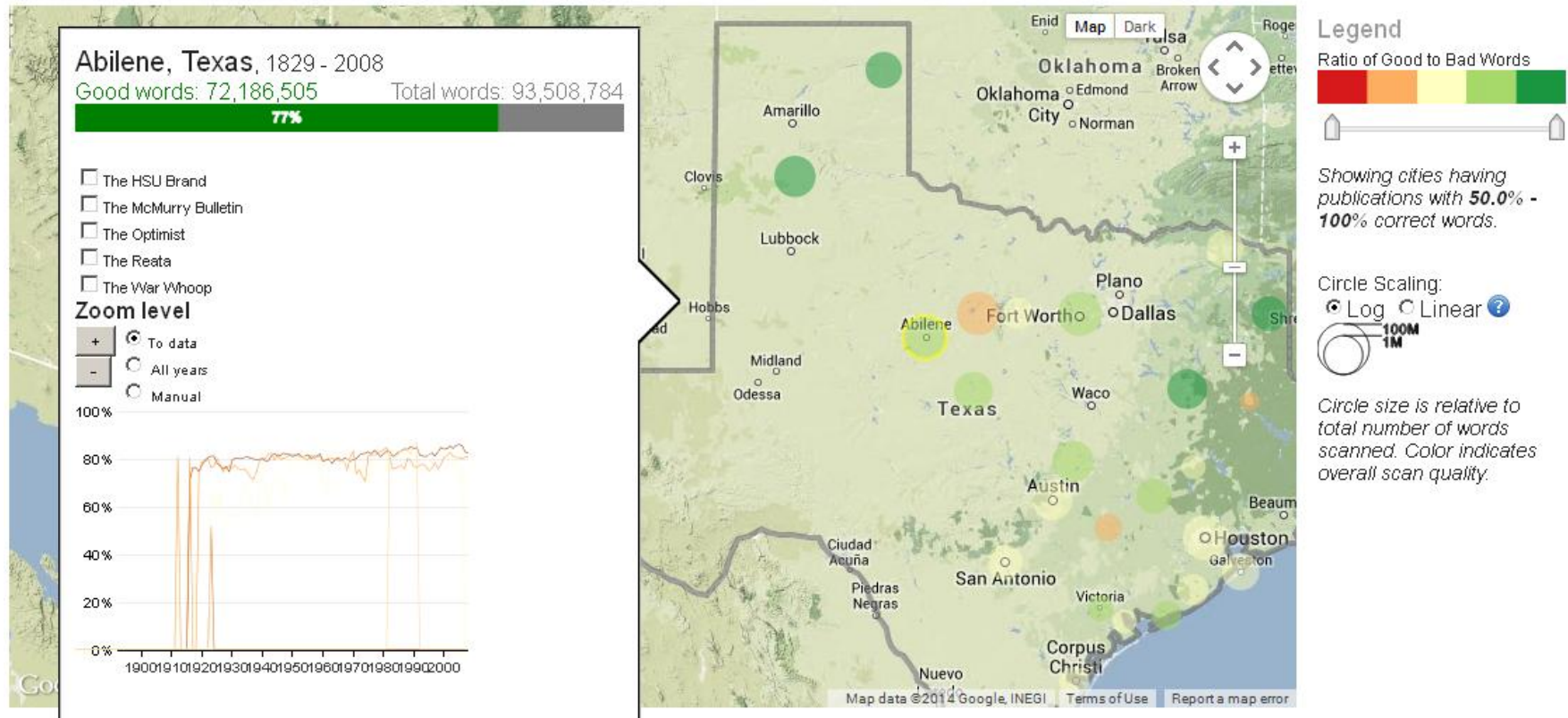Loading Complete

Filter publications...

▼ By language

☐ All Languages

- ☐ English
- ☐ Spanish
- ☐ German
- ☐ Chinese
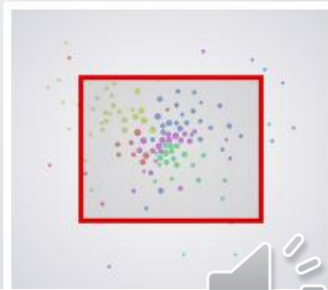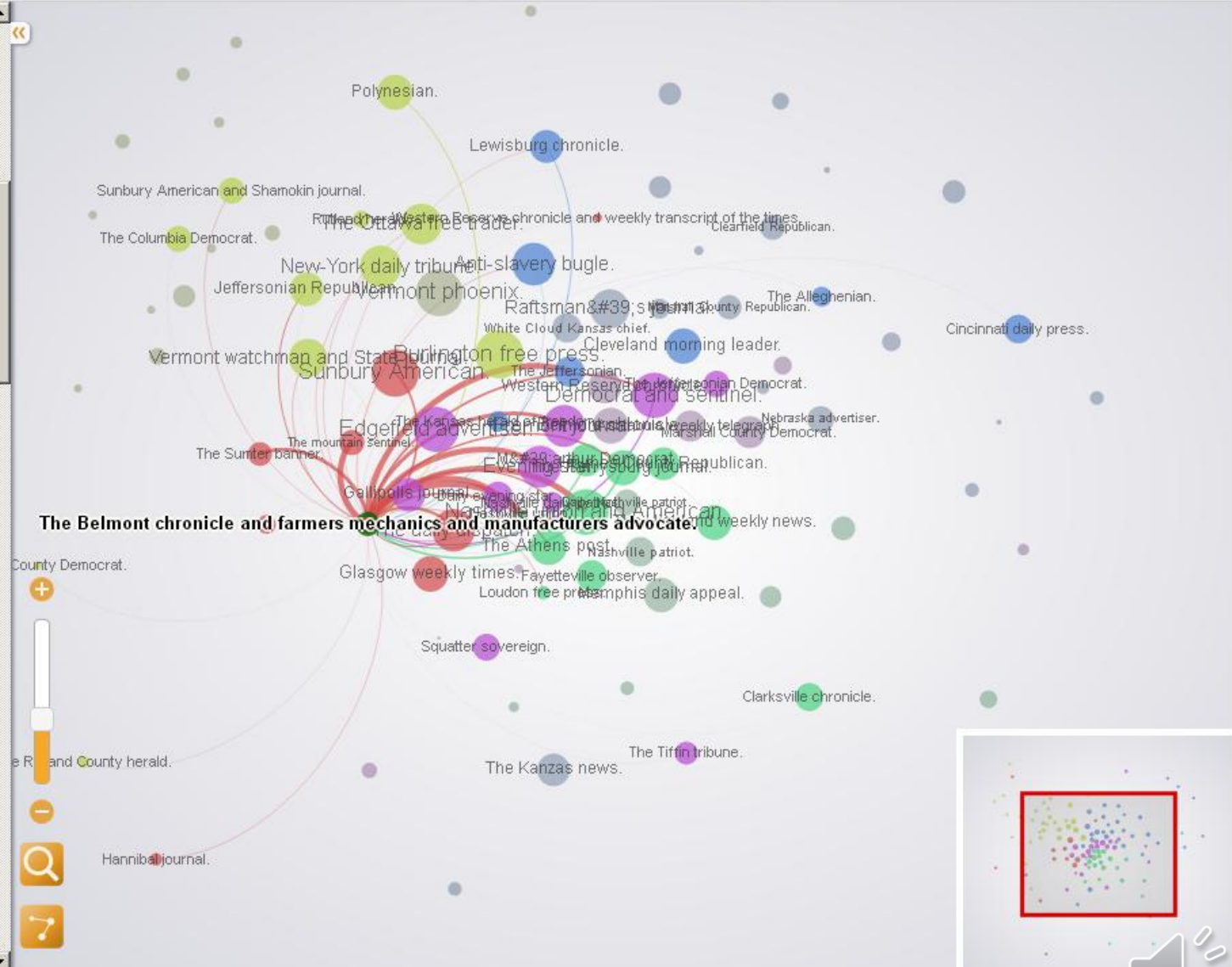- ☐ Swedish
- ☐ Italian
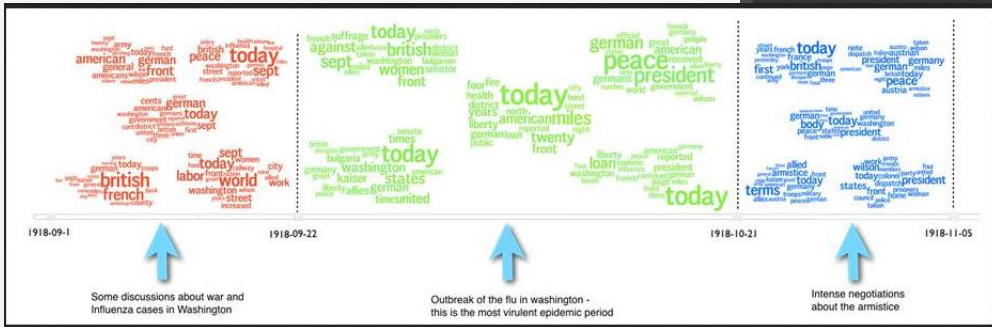- ☐ French
- ☐ Other

▶ By publication frequency

1918-09-1   1918-09-22   1918-10-21   1918-11-05

Some discussions about war and Influenza cases in Washington

Outbreak of the flu in washington - this is the most virulent epidemic period

Intense negotiations about the armistice

Space Collection Quantity and Quality by Location

Abilene, Texas, 1829 - 2008

Good words: 72,186,505   Total words: 93,508,784

77%

- ☐ The HSU Brand
- ☐ The McMurry Bulletin
- ☐ The Optimist
- ☐ The Reata
- ☐ The War Whoop

Zoom level

- ○ To data
- ○ All years
- ○ Manual

Legend

Ratio of Good to Bad Words

Showing cities having publications with **50.0**% - **100**% correct words.

Circle Scaling:
- ○ Log ○ Linear ❓

100M
1M

Circle size is relative to total number of words scanned. Color indicates overall scan quality.

Gp Gephi

Sunbury American.

The Jeffersonian Democrat. [3]
The Emporia news.
The Kansas herald of freedom. [2]
The Kanzas news. [7]
The Belmont chronicle and farmers mechanics and manufacturers advocate.
Gallipolis journal. [12]
Holmes County farmer.
Vermont watchman and State journal. [14]
The Illinois free trader and LaSalle County commercial advertiser.
Burlington free press. [18]
Fayetteville observer. [5]
Edgefield advertiser. [19]
The Ottawa free trader. [10]
Memphis daily appeal. [10]
M'arthur Democrat. [12]
The daily dispatch. [24]
Squatter sovereign. [9]
The Columbia Democrat. [7]
The Illinois free trader.
Anti-slavery bugle. [14]
The Somerset herald and farmers' and mechanics' register. [2]
Lewisburg chronicle. [31]
The Northern galaxy.
Columbia Democrat and Bloomsburg general advertiser.
Polynesian. [11]
Marshall County Republican. [4]
Sunbury American and Shamokin journal. [6]
The Jeffersonian. [11]
The examiner. [2]
The Middlebury register. [3]
Marshall County Democrat. [3]
Clearfield Republican. [6]
Nebraska advertiser.
Nashville union. [7]
The Sumter banner. [6]
Loudon free press. [2]

Polynesian.
Lewisburg chronicle.
Sunbury American and Shamokin journal.
The Columbia Democrat.
Ripley Western Reserve chronicle and weekly transcript of the times.
Clearfield Republican.
The Ottawa free trader.
New-York daily tribune.
Anti-slavery bugle.
Jeffersonian Republican.
Vermont phoenix.
The Alleghenian.
Raftsman's journal.
Marshall County Republican.
White Cloud Kansas chief.
Cincinnati daily press.
Cleveland morning leader.
Vermont watchman and State
Burlington free press.
Sunbury American.
The Jeffersonian.
Democrat and sentinel.
Western Reserve
The Jeffersonian Democrat.
Nebraska advertiser.
Edgefield advertiser.
Marshall County Democrat.
The mountain sentinel.
The Sumter banner.
Evening star
Gettysburg journal.
Gallipolis journal.
Nashville union and American.
weekly news.
The Athens post.
Nashville patriot.
Glasgow weekly times.
Fayetteville observer.
Loudon free press.
Memphis daily appeal.
The Belmont chronicle and farmers mechanics and manufacturers advocate.
County Democrat.
Squatter sovereign.
Clarksville chronicle.
County herald.
The Tiffin tribune.
The Kanzas news.
Hannibal journal.

How did we get from this⋯

… to this?!

The Growth of Newspapers Across the U.S.: 1690-2011

1945
World War II
14083 listings

World War II sapped rural newsrooms of manpower as young men enlisted in droves after the Pearl Harbor attacks in 1941. As in many other industries, women filled the gap in large numbers, and legal protections were passed for newspapers that needed to suspend publication during the War. Newspapers placed new emphasis on maps and information graphics to help explain the complex and far flung battlefields to the public. Back home, the War Relocation Authority permitted Japanese-Americans to publish their own newspapers in detention camps like Manzanar in California and Gila River in Arizona.

National Digital Newspaper Program Partners:
36 institutions | 7 million pages now online | 1836-1922

National Digital Newspaper Program, 2005-2013

National Digital Newspaper Program, 2005-

GOALS:

- To enhance access to historic American newspapers from every state and territory

- To develop best practices for the digitization of historic newspapers (shared community)

- Free and open content, available to all

NDNP: Built for Sustainability

- Shared technical specifications
- Institutional cooperation
- Data integrity
- Data management
- Expect change

Shared technical specifications

NDNP Technical Guidelines for Applicants

(NDNP Tech Specs)

http://www.loc.gov/ndnp/guidelines/

Shared technical specifications

- ## Stable guidelines since 2005

- ## Changes limited to:
  - ### Clarifications to existing practice
  - ### Version updates (ALTO 2.0)
  - ### Expanding content scope
  - ### Simplifying some metadata requirements
    - #### Technical metadata from Microfilm collation now OPTIONAL
    - #### Section / Edition Labels now OPTIONAL

The National Digital Newspaper Program (NDNP)
Technical Guidelines for Applicants

Document Control

Revision	Date	Summary of Change

Overview of Technical Approach for 2012-14 NDNP Awards

The National Digital Newspaper Program is a long-term effort and the technical environment will change as the program continues. The National Endowment for the Humanities (NEH) and the Library of Congress (LC) have selected a technical approach to balance long-term objectives and shorter-term constraints. These include:

- convenient accessibility over the World Wide Web for the general public to the entire collection as it grows, through a consistent interface and using proven technology;
- page images of sufficient spatial and tonal resolution to support effective performance of OCR (optical character recognition) software and representation of printed half-tones, given the limitations of microfilm, expecting that future improvements in OCR and image processing will be applied to the same images;
- the use of digital formats with a high probability of sustainability - in particular, using standard formats where possible and proprietary formats only where widely adopted;
- and attention to the cost of digital conversion and maintenance of the resulting assets.

The goal of the initial program phase is to enhance access to newspapers through a Web-accessible NDNP delivery application, open and freely available to the public. This delivery application will provide access to a substantive amount of newspaper content selected for its historical value with broad geographic coverage and consistent sustainable

Library of Congress
09/20/2011
Page 2

Distributed Project / Shared Effort

NDNP Data

For every page:

- Archival Image: TIFF
- Production Image: JPEG 2000
- Printable Image: PDF
- OCR XML File: ALTO

For every issue and reel:

- METS XML File
  - Descriptive metadata
  - Structural metadata
  - Preservation metadata

For newspaper title:

- MARC record
  - Geographic metadata
  - Time Period
  - Subject metadata

http://www.loc.gov/ndnp/guidelines/

METS ALTO

- NDNP ALTO specification requires:
  - Column level text block zoning and
  - Coordinates to map or highlight text to image files.

- Chronicling America supports page level access with visual representation of search results.

Page Level Access

Zoom, pan, and clip tools are available

Chronicling America - LC

Chronam: Open Source software

https://github.com/LibraryofCongress/chronam

chronam

http://nyshistoricnewspapers.org/

http://oregonnews.uoregon.edu/

Chronicling America - API

http://chroniclingamerica.loc.gov/about/api/

Chronicling America – Bulk Data: OCR Downloads for External Services

OCR Data

What is OCR?

Optical character recognition (OCR) is a fully automated process that converts the visual image of numbers and letters into computer-readable numbers and letters. Computer software can then search the OCR-generated text for words, phrases, numbers, or other characters. However, OCR is not 100 percent accurate, and, particularly if the original item has extraneous markings on the page, unusual text styles, or very small fonts, the searchable text OCR generates will contain errors that cannot be corrected by automated means.

Although errors in the process are unavoidable, OCR is still a powerful tool for making text-based items accessible to searching. For example, important concept words often appear more than once within an article. Therefore, if OCR misreads one instance of a key word in a passage, but correctly reads the second instance, the passage will still be found in a full-text search.

To enable research and external services Chronicling America provides bulk access to its OCR data. The table below itemizes a list of data files available for download. Each file will decompress into directory structure that lets you easily map the OCR file to the URL identifier for that page. For example a file such as sn830030214/1903/05/01/ed-1/seq-1/ocr.txt maps to the URL http://chroniclingamerica.loc.gov/lccn/sn830030214/1903-05-01/ed-1/seq-1/.

If you are interested in automated access to this data you may want to use the Atom and JSON versions of this table.

Filename	Batch	Created	Size	SHA-1 Checksum
batch_wa_indianplum_ver02.tar.bz2	batch_wa_indianplum_ver02	2014-01-14T21:54:30-05:00	2.6 GB	b7131b22dc5f2e22af04c77ea5471db782a94f63
batch_wvu_belgium_ver01.tar.bz2	batch_wvu_belgium_ver01	2014-01-14T21:51:48-05:00	2.4 GB	819971e355c08cd032742891dd051a62bf5c4529
batch_wvu_denmark_ver01.tar.bz2	batch_wvu_denmark_ver01	2014-01-14T21:44:51-05:00	2.1 GB	627346b3b69e31a59a1d6684a024fe08d2e0ff17
batch_wvu_eritrea_ver01.tar.bz2	batch_wvu_eritrea_ver01	2014-01-14T21:31:30-05:00	2.1 GB	208fb744dd59f63efb59891de3b14e6bb3d60b75
batch_vtu_graniteville_ver01.tar.bz2	batch_vtu_graniteville_ver01	2014-01-14T21:20:57-05:00	1.6 GB	6e3aff822b8e1b640e8624a164f3965889069b04
batch_vtu_burlington_ver01.tar.bz2	batch_vtu_burlington_ver01	2014-01-14T21:16:31-05:00	1.7 GB	f41c1c8117d77ebf7b760300b90da457b218d7d7
batch_wa_fir_ver01.tar.bz2	batch_wa_fir_ver01	2014-01-14T21:09:18-05:00	1.7 GB	d94d31a3d548151bc7440dadf2d291ac7213f475
batch_wvu_austria_ver01.tar.bz2	batch_wvu_austria_ver01	2014-01-14T21:06:36-05:00	1.7 GB	e4fb071412b629a190d732fc3d400937a5b8f0d6

http://chroniclingamerica.loc.gov/ocr/
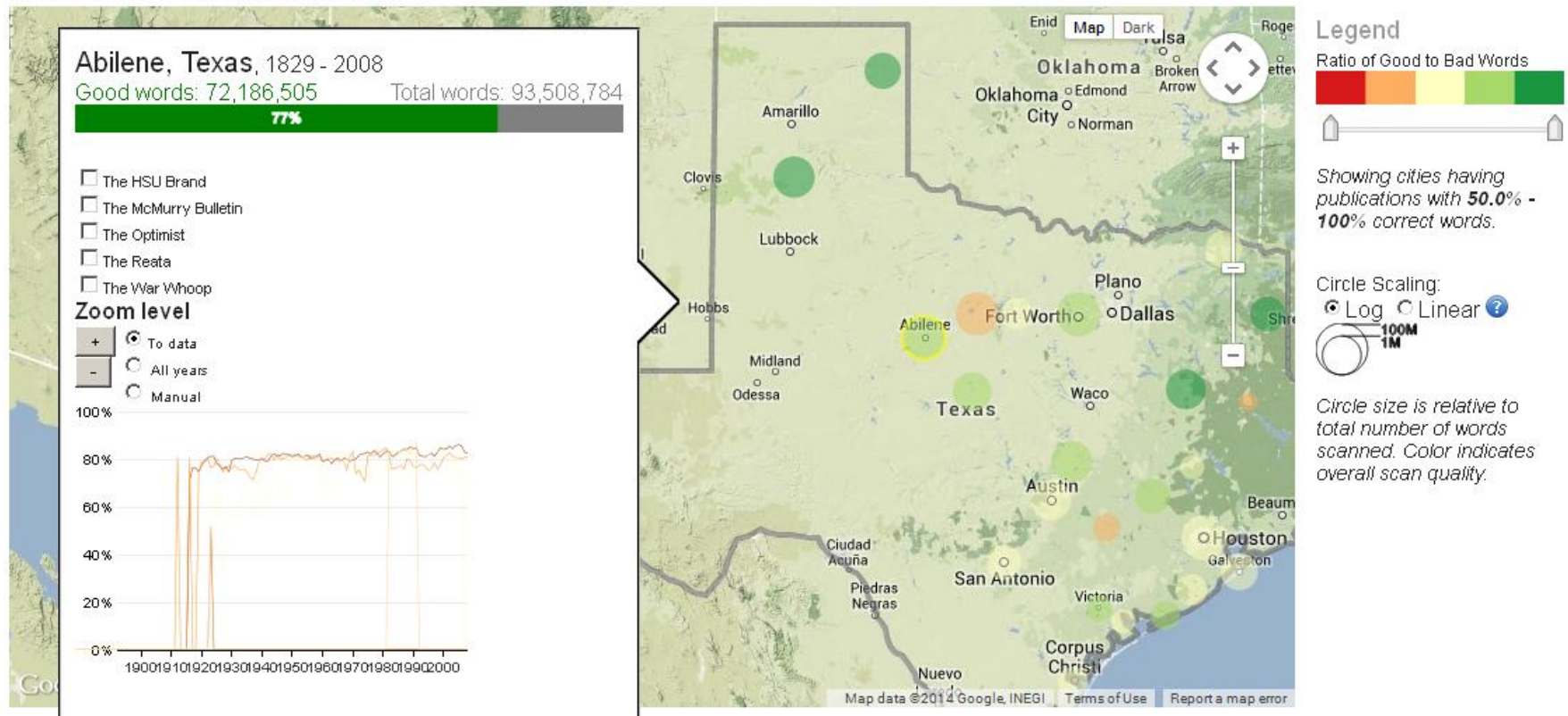
Advancing Research with Chronam
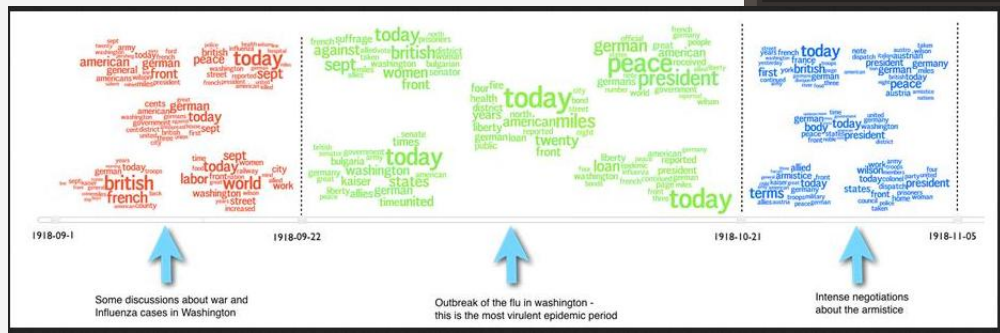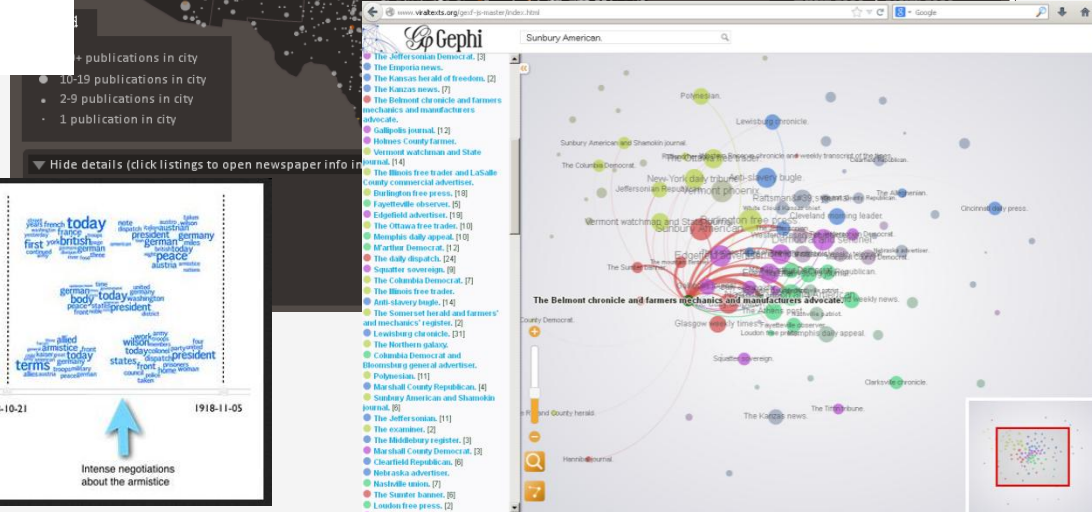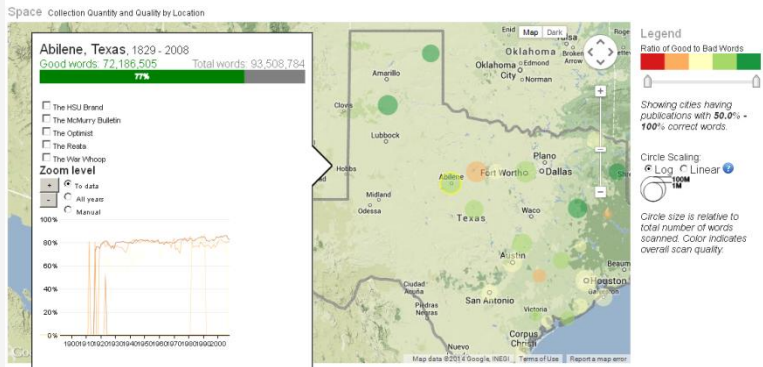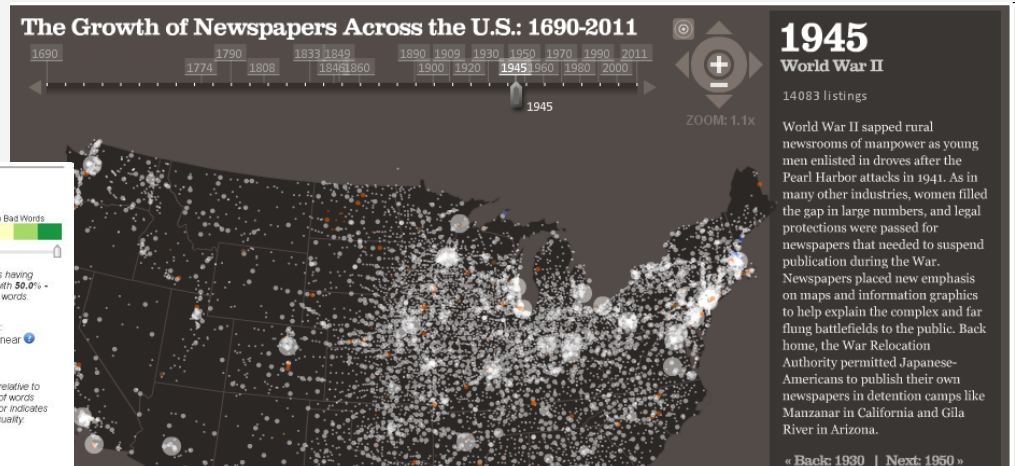
The Growth of US Newspapers, 1690-2011
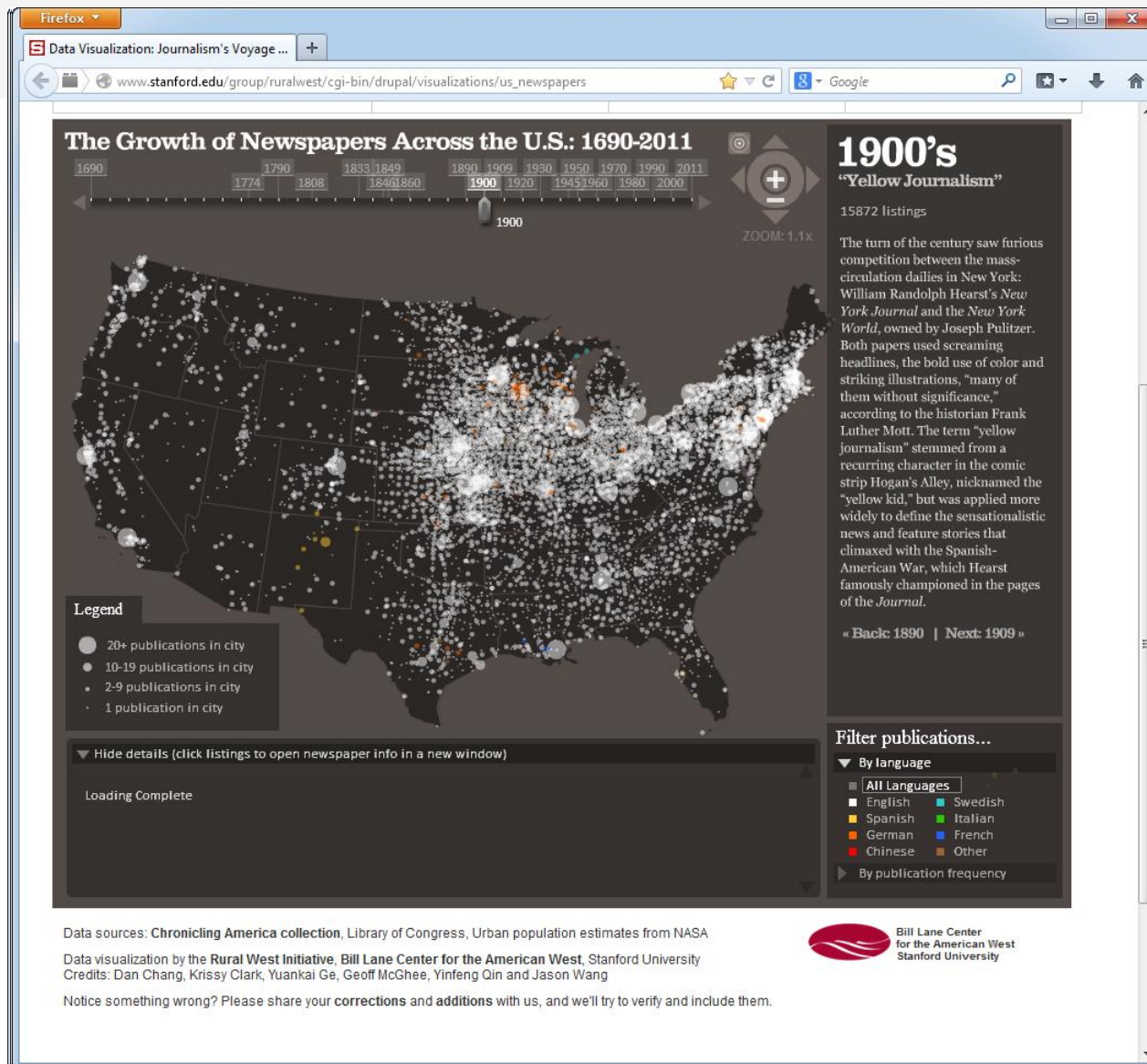
Mapping Texts

Mapping Texts is a collaboration between scholars, staff and students at Stanford University and the University of North Texas. It is supported by the National Endowment for the Humanities.
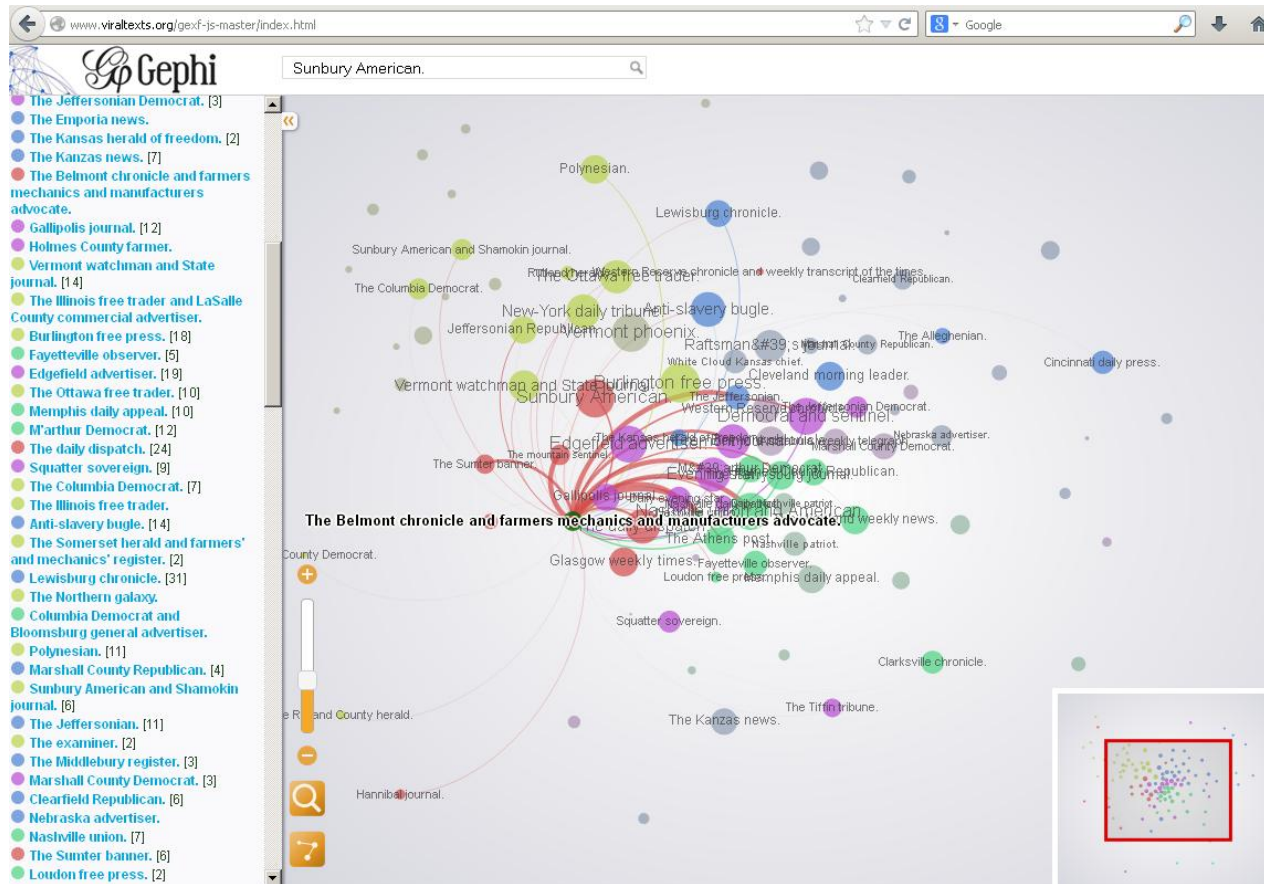
http://mappingtexts.org/

"An Epidemiology of Information: Data Mining the 1918 Influenza Pandemic"
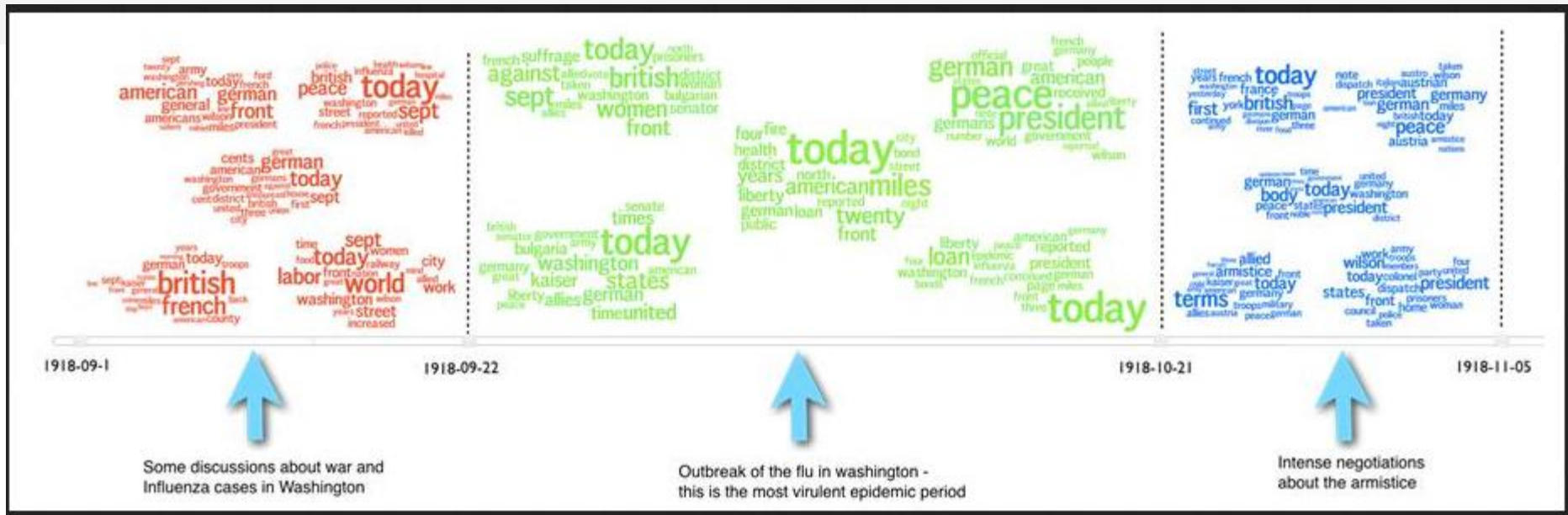
"The goal … is to develop methods for combining algorithmic techniques with the interpretive strength of traditional historical and rhetorical analysis in order to help researchers better understand reporting on the 1918 flu pandemic in American and Canadian newspapers."

- Text mining
- Topic Modeling
- Tone Analysis

Professor E. Thomas Ewing, Principal Investigator and Project Director, Department of History, Virginia Tech

http://www.flu1918.lib.vt.edu/

Viral Networks in 19th-Century Newspapers

"*Infectious Texts* is sponsored by Northeastern University's NULab for Texts, Maps, and Networks and generously funded by the National Endowment for the Humanities' Office of Digital Humanities. The project team includes Professors Ryan Cordell, Elizabeth Maddock Dillon, and David Smith, as well as Ph.D. students Abby Mullen and Matthew Williamson."

http://www.viraltexts.org/

Chronicling America

Weathering the Storm

http://chroniclingamerica.loc.gov/lccn/sn83045433/1912-11-30/ed-1/seq-1 /

Beyond the *Chronicling America* Web Site

Regular update/highlights information by RSS* or email subscription

- http://www.loc.gov/rss/ndnp/ndnp.xml; and **Recent Additions RSS feed** (newspaper titles added)
- Keep up with what's in *Chronicling America*!

Open-source code for chronam application available on GitHub

- https://github.com/LibraryOfCongress/chronam
- chronam is the Django application that the Library of Congress uses to make its
- Chronicling America website (a core set of functionality for loading, modeling and
- indexing NDNP data)
- Check it out and take it for a spin!

Linked data and API** access in *Chronicling America*

- RDF/XML views available to Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
- OpenSearch and Atom APIs for search queries, Bookmarkable URIs for all pages
- http://chroniclingamerica.loc.gov/about/api/

NDNP Extras! from the NDNP program site

- http://www.loc.gov/ndnp/extras/
- Visualizations, tutorials, podcasts, teaching resources, state project blogs, how Chronicling America is used in other projects and more!

Thank you!

- **NDNP Public Web**
  http://www.loc.gov/ndnp/

- **NDNP Web Service**

  Chronicling America: Historic American Newspapers
  http://chroniclingamerica.loc.gov

- **Contact us at ndnptech@loc.gov**