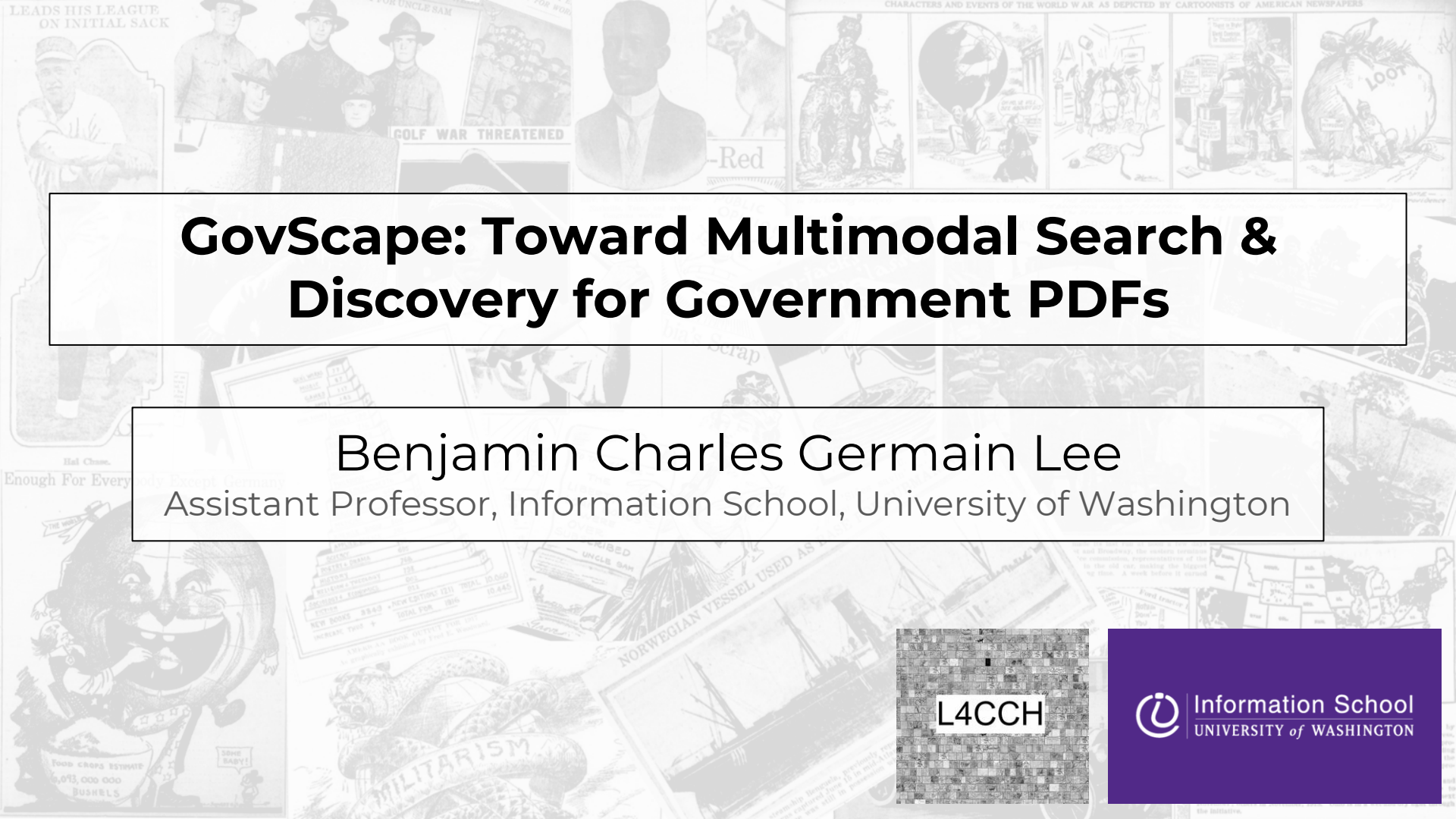




# GovScape: Toward Multimodal Search & Discovery for Government PDFs

Benjamin Charles Germain Lee

Assistant Professor, Information School, University of Washington

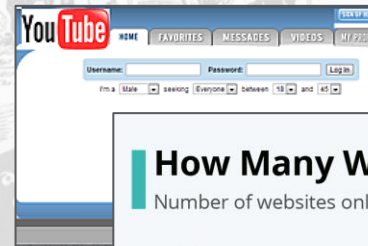
L4CCH



Information School  
UNIVERSITY of WASHINGTON

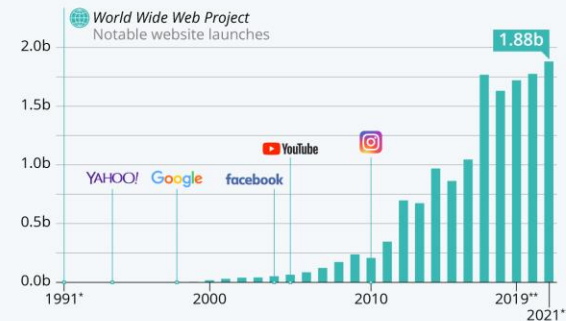


Library of Congress Serials



## How Many Websites Are There?

Number of websites online from 1991 to 2021



\* As of August 1, 1991.

\*\* Latest available data for 2019: October 28, for 2020: June 2, for 2021: August 6.

Source: Internet Live Stats



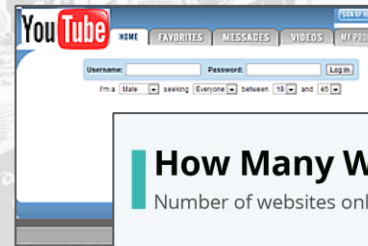
statista

The World Wide Web



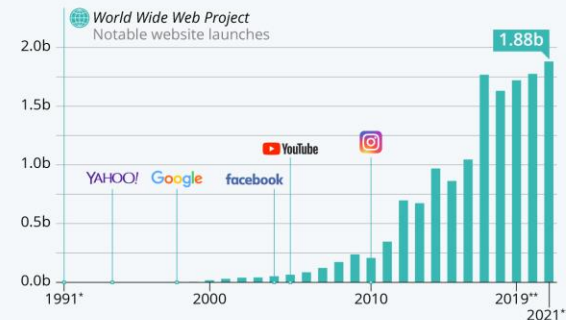
20+ million  
pages  
(100+ terabytes)

Library of Congress Serials



## How Many Websites Are There?

Number of websites online from 1991 to 2021



\* As of August 1, 1991.

\*\* Latest available data for 2019: October 28, for 2020: June 2, for 2021: August 6.

Source: Internet Live Stats



statista

The World Wide Web



**20+ million  
pages  
(100+ terabytes)**

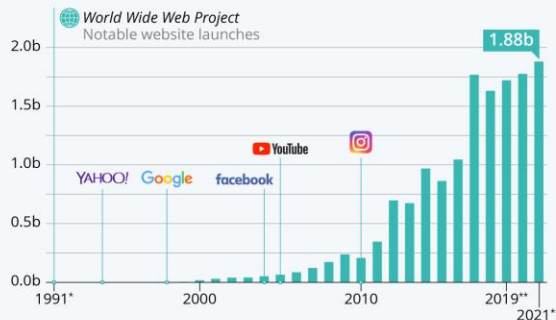
Library of Congress Serials



**Billions of web  
page snapshots  
(petabyte-scale)**

## How Many Websites Are There?

Number of websites online from 1991 to 2021



\* As of August 1, 1991.

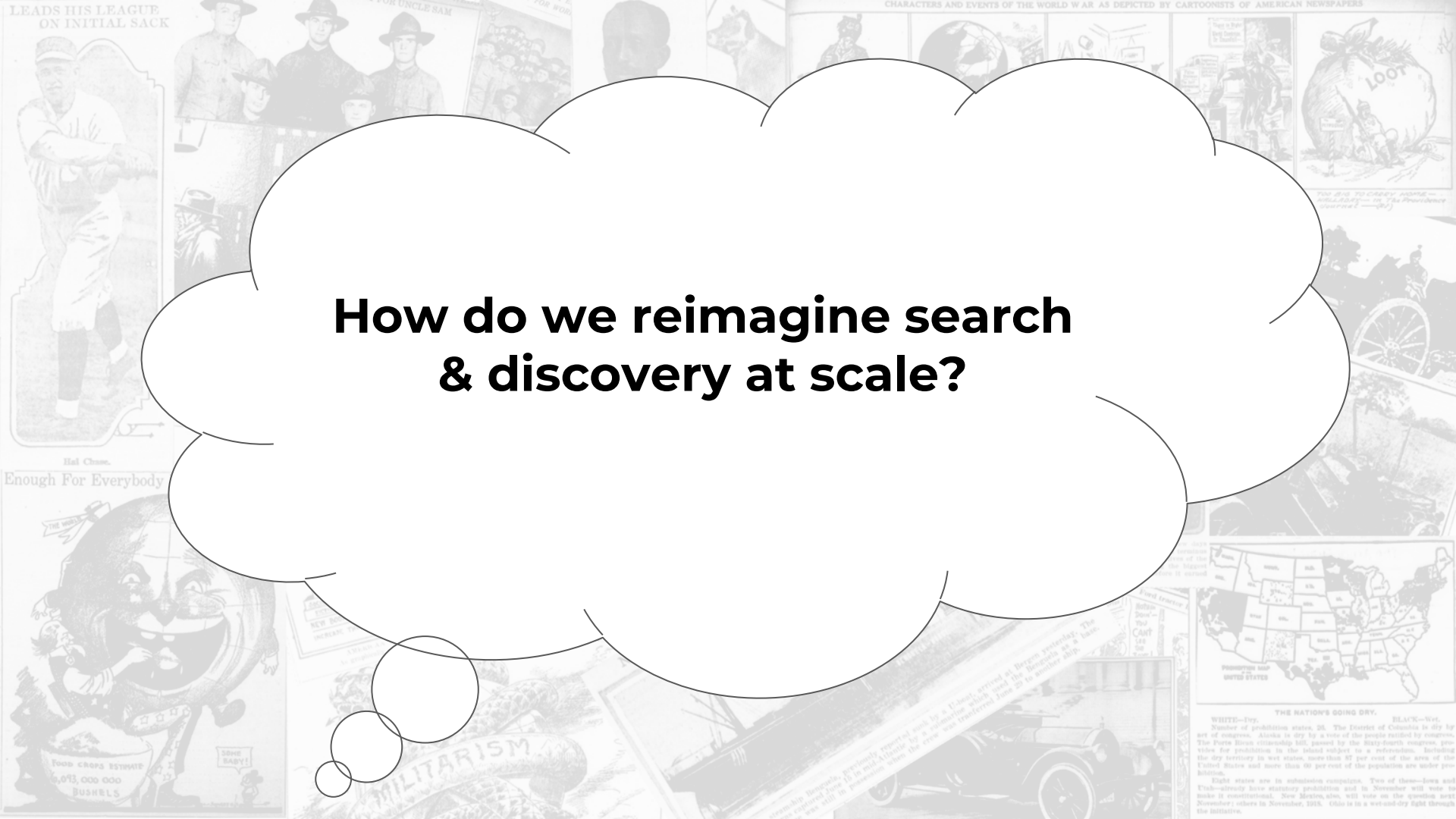
\*\* Latest available data for 2019: October 28, for 2020: June 2, for 2021: August 6.

Source: Internet Live Stats



statista

The World Wide Web



# How do we reimagine search & discovery at scale?

# Computing Cultural Heritage

# Computing Cultural Heritage

1. Building **large-scale search and discovery systems** for digital collections using AI.

# Computing Cultural Heritage

1. Building **large-scale search and discovery systems** for digital collections using AI.
2. Exploring how a **wide range of patrons**, from scholars to members of the public, can interact with these systems in order to enrich their experiences.

# My Academic Home



Information School  
UNIVERSITY of WASHINGTON

PAUL G.  
ALLEN  
SCHOOL  
W

Lab for Computing  
Cultural Heritage

# Libraries, Archives, and Museums

LIBRARY  
OF CONGRESS  
**KLUGE**

LIBRARY  
OF CONGRESS  
**LABS**

UNITED STATES  
**HOLOCAUST**  
MEMORIAL  
**MUSEUM**

**D P L A**  
DIGITAL PUBLIC LIBRARY OF AMERICA

Lab for Computing  
Cultural Heritage

**Library  
Innovation  
Lab**

LEADS HIS LEAGUE ON INITIAL SACK

UNCLE SAM

GOLF WAR THREATENED

Red

REV. E. W. HARTSHORN, D. D.

PUBLIC OPINION

flap-jacks encircles

NEW YORK'S LAST HORSE CAR QUILTS

What does computing cultural heritage mean in practice?

LOOT

THE NATION'S GOING DRY.

WHITE—Dry. Number of prohibition states, 25. The States of Columbia is dry by act of congress. Alaska is dry by a vote of the people ratified by congress. The Prohibition Congress, passed by the Sixty-ninth congress, provides for prohibition in the third subject to a referendum. Including the dry territory in wet states, more than 92 per cent of the area of the United States and more than 60 per cent of the population are under prohibition.

BLACK—Wet. Eight states are in prohibition companies. Two of these—long and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico will vote on the question next November; others in November, 1915. This is a wet-and-dry fight through the initiative.

LOAN BONDS WERE OVER SUBSCRIBED

NORWEGIAN VESSEL USED AS BASE FOR S

MILITARISM

THE NATION'S GOING DRY.

WHITE—Dry. Number of prohibition states, 25. The States of Columbia is dry by act of congress. Alaska is dry by a vote of the people ratified by congress. The Prohibition Congress, passed by the Sixty-ninth congress, provides for prohibition in the third subject to a referendum. Including the dry territory in wet states, more than 92 per cent of the area of the United States and more than 60 per cent of the population are under prohibition.

BLACK—Wet. Eight states are in prohibition companies. Two of these—long and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico will vote on the question next November; others in November, 1915. This is a wet-and-dry fight through the initiative.

# The International Tracing Service Archive's Central Name Index



|                                      |              |                       |
|--------------------------------------|--------------|-----------------------|
| SCHORR Rita                          |              | 290 683               |
| 16.4.28                              | Borislau     | fr. Pole<br>rel. jüd. |
| 7.41                                 | Gh. Borislau |                       |
| 11.41                                | ZAL "        |                       |
| 4.44                                 | KZ Flassow   |                       |
| 8.44                                 | KZ Auschwitz |                       |
| 2.45                                 | " befr.      |                       |
| BLR, München<br>11/6-1968-Har-Tepfer |              |                       |

|                       |                                            |           |        |
|-----------------------|--------------------------------------------|-----------|--------|
| Date                  | 21.12.50/H.Gh                              | File #    | 18-151 |
| Name                  | SCHORR Rita                                | Nat       | Jewish |
| BD                    | 1908 BP Drohobycz                          |           |        |
| Next of Kin           |                                            |           |        |
| Source of Information | Jew. Comm. Münch., thr. AJDC Bergen/Belsen |           |        |
| Last kn. Location     | München, Keimtenbadstr. 22/Date            | 1945-1948 |        |
| CC/Reason             | Aut.                                       | lib.      |        |
| Transf. on            | to                                         |           |        |
| Died on               | in                                         |           |        |
| Cause of death        |                                            |           |        |
| Buried on             | in                                         |           |        |
| Grave                 |                                            | D.C. No.  |        |
| Remarks               | Nr. 5668                                   |           |        |

|                                   |                                                   |
|-----------------------------------|---------------------------------------------------|
| Häftl. Nr.:                       |                                                   |
| Name:                             | SCHORR Rita                                       |
| geb.                              |                                                   |
| am:                               | in:                                               |
| Weitere Angaben<br>siehe Original | Sachdok.-Ordner: 6<br>V.M. Gal'zien<br>Seite: 349 |

(40 million cards)

# Newspaper Navigator

Library of Congress's  
first in-house AI  
application

Search My Collection Train My AI Navigators Chronicling America LC Labs

Save

Select an AI navigator:

baseball players

+ New AI Navigator

(+) The AI navigator will try to find photos like the ones selected below!



(-) The AI navigator will try to avoid photos like the ones selected below!



baseball players

Below, your AI navigator has found visually similar photos based on the selections on the left. You can re-train your navigator by adjusting your selections on the left and making new selections below.



PAUL G.  
ALLEN  
SCHOOL  
W

LIBRARY  
OF CONGRESS  
LABS

NATIONAL DIGITAL  
NEWSPAPER PROGRAM

news-navigator.labs.loc.gov

**THE EVENING STAR**  
 PUBLISHED DAILY  
 BY THE STAR PUBLISHING CO.  
 1000 K STREET, N.W.  
 WASHINGTON, D. C. 20004  
 Phone: 223-1234

# The Evening Star

**Friday, June 10, 1966**  
 Edition: Final  
 Price: 10¢

## NAZI COUNTERATTACKS REPULSED NEAR CAEN

**Allied Military Forces Pushed Back Into North of France**

**Allies Clear Enemy From Landing Areas Battle Raging Inland**

General Eisenhower's forces pushed back into the north of France after a series of heavy German counterattacks near Caen, France.

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**Allies Regain North of France**

The German forces, which had been pushing back into the Allied positions, were repulsed after a series of heavy battles. The Allied forces, which had been pushing back into the German positions, were repulsed after a series of heavy battles.

**THE EVENING STAR**  
 PUBLISHED DAILY  
 BY THE STAR PUBLISHING CO.  
 1000 K STREET, N.W.  
 WASHINGTON, D. C. 20004  
 Phone: 223-1234

# The Evening Star

**Friday, June 10, 1966**  
 Edition: Final  
 Price: 10¢

## NAZI COUNTERATTACKS REPULSED NEAR CAEN

**Allied Infantry  
 Pushes Off  
 North of Rome**

**Allies Clear Enemy  
 From Landing Areas  
 Battle Raging Inland**

General Lucian K. Truscott's 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy.

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**Allies Repulse  
 100 German Tanks  
 Near Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**100 German Tanks  
 Repulsed Near  
 Caen**

The 1st Infantry Division is pushing off the German counterattacks north of Rome, Italy. The division is pushing off the German counterattacks north of Rome, Italy.

**Special Advertising Section**  
 Classified Advertising  
 Classified Advertising  
 Classified Advertising

# The Evening Star

**Friday, June 10, 1944**  
 Edition 1  
 Price 10 Cents

## NAZI COUNTERATTACKS REPULSED NEAR CAEN

**Allied Military Forces**  
 Pushed Back Into North of River

German Losses  
 Heavy, According to  
 U.S. Forces

**Allies Clear Enemy From Landing Areas**  
 Battle Raging Inland  
 German Forces  
 50 Miles, Back From  
 Seals in Bay of Normandy

**U.S. Forces Repulse**  
 All German Counter  
 Attacks Near Caen  
 German Forces  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**U.S. Forces**  
 Pushed Back Into  
 North of River

**THE EVENING STAR**  
 PUBLISHED DAILY  
 BY THE STAR PUBLISHING CO.  
 1000 K STREET, N.W.  
 WASHINGTON, D. C. 20004  
 Phone: 223-1234

# The Evening Star

**Friday, June 10, 1966**  
 Edition: Final  
 Price: 10¢

## NAZI COUNTERATTACKS REPULSED NEAR CAEN

**Allied Military Forces Pushed Back Into North of France**

**Allies Clear Enemy From Landing Areas Battle Raging Inland**

General Eisenhower's forces have pushed back into the north of France after a series of heavy Nazi counterattacks near Caen, France.

The fighting was the heaviest since the Allied invasion of Normandy on June 6. The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland.

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Allies Fighting Off Enemy Counterattacks**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

**Heavy Fighting Continues**

The Allies have now cleared the enemy from the landing areas and are now fighting a battle of attrition inland. The fighting was the heaviest since the Allied invasion of Normandy on June 6.

[illegible][illegible]

**Comic Strip Panels 1-4:**

Panel 1: A man in a suit and hat is talking to a woman in a dress. He says, "I'm going to the bank to get some money." She replies, "That's all right, but don't forget to get some more for the kids." Panel 2: The man is at the bank, talking to a teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 3: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids." Panel 4: The man is at the bank again, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?"

**Comic Strip Panels 5-8:**

Panel 5: The man is at the bank, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 6: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids." Panel 7: The man is at the bank again, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 8: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids."

**Comic Strip Panels 9-12:**

Panel 9: The man is at the bank, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 10: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids." Panel 11: The man is at the bank again, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 12: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids."

**Comic Strip Panels 13-16:**

Panel 13: The man is at the bank, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 14: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids." Panel 15: The man is at the bank again, talking to the teller. He says, "I'd like to get some money, please." The teller replies, "Sure, how much?" Panel 16: The man is back home, talking to the woman. He says, "I got the money, but the teller said I had to wait for an hour." She replies, "That's all right, but don't forget to get some more for the kids."

**Local, Foreign and Foreign**

**Local:** The local section contains various news items, including a report on a fire at a local business, a notice about a lost dog, and a mention of a local sports event.

**Foreign:** The foreign section includes news from various international locations, such as a report on a political event in Europe and a notice about a cultural festival in Asia.

**Foreign:** This section continues with more international news, including a report on a natural disaster in South America and a notice about a diplomatic meeting in Africa.

**Local, Foreign and Foreign**

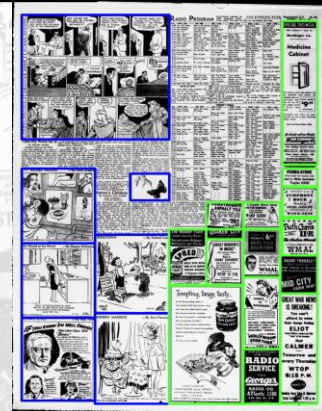
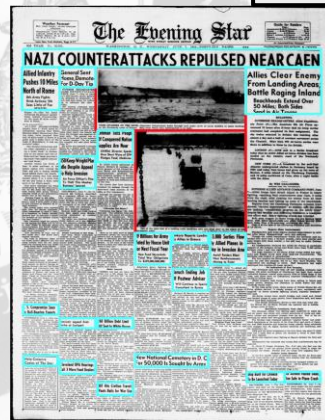
**Local:** The local section contains various news items, including a report on a fire at a local business, a notice about a lost dog, and a mention of a local sports event.

**Foreign:** The foreign section includes news from various international locations, such as a report on a political event in Europe and a notice about a cultural festival in Asia.

**Foreign:** This section continues with more international news, including a report on a natural disaster in South America and a notice about a diplomatic meeting in Africa.

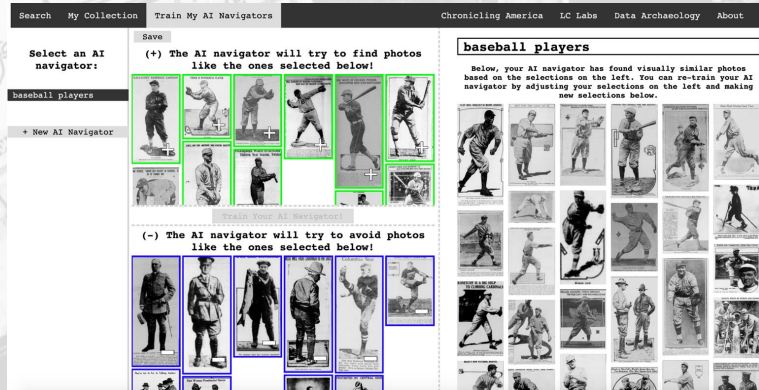
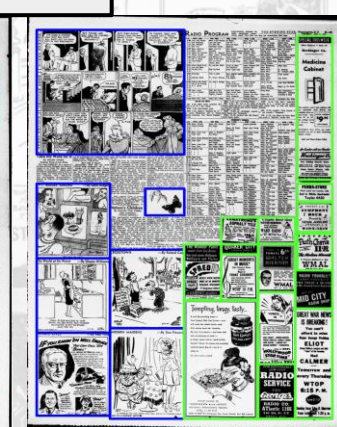
ATED BY CARTOONISTS OF AMERICAN NEWSPAPERS

## Dataset



| Income Range            | Number of People |
|-------------------------|------------------|
| Over \$1,000,000        | 1                |
| \$750,000 - \$1,000,000 | 1                |
| \$500,000 - \$750,000   | 1                |
| \$250,000 - \$500,000   | 1                |
| \$100,000 - \$250,000   | 1                |
| \$75,000 - \$100,000    | 1                |
| \$50,000 - \$75,000     | 1                |
| \$25,000 - \$50,000     | 1                |
| \$10,000 - \$25,000     | 1                |
| \$5,000 - \$10,000      | 1                |
| \$2,500 - \$5,000       | 1                |
| \$1,000 - \$2,500       | 1                |
| Under \$1,000           | 1                |

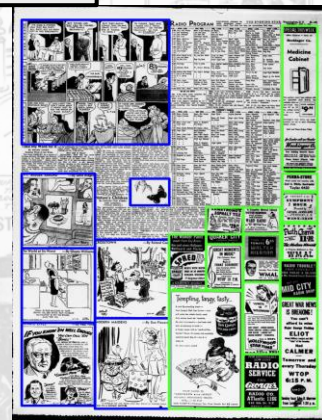
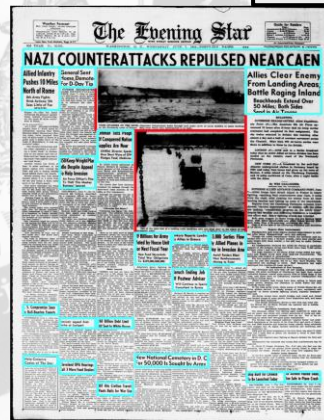
Source: U.S. Census Bureau, 1992



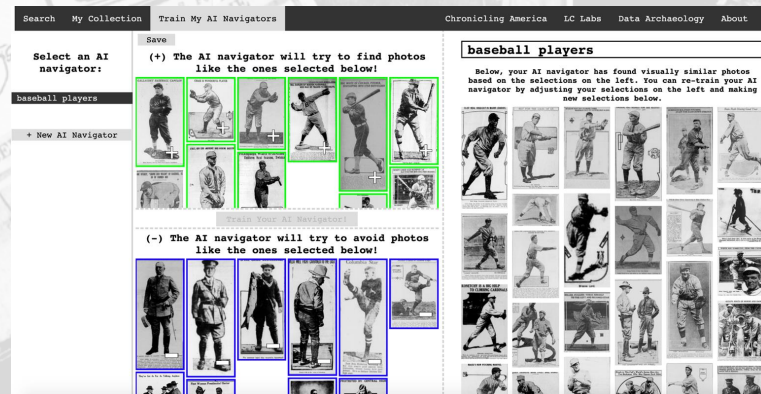
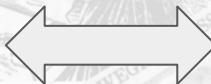
Eight states are in submission campaigns. Two of these—Louisiana and Utah—already have statutory prohibition and in November will vote to make it constitutional. New Mexico, also, will vote on the question next November; others in November, 1918. Ohio is in a wet-and-dry fight through the initiative.

LEADS HIS LEAGUE  
ON INITIAL SACK

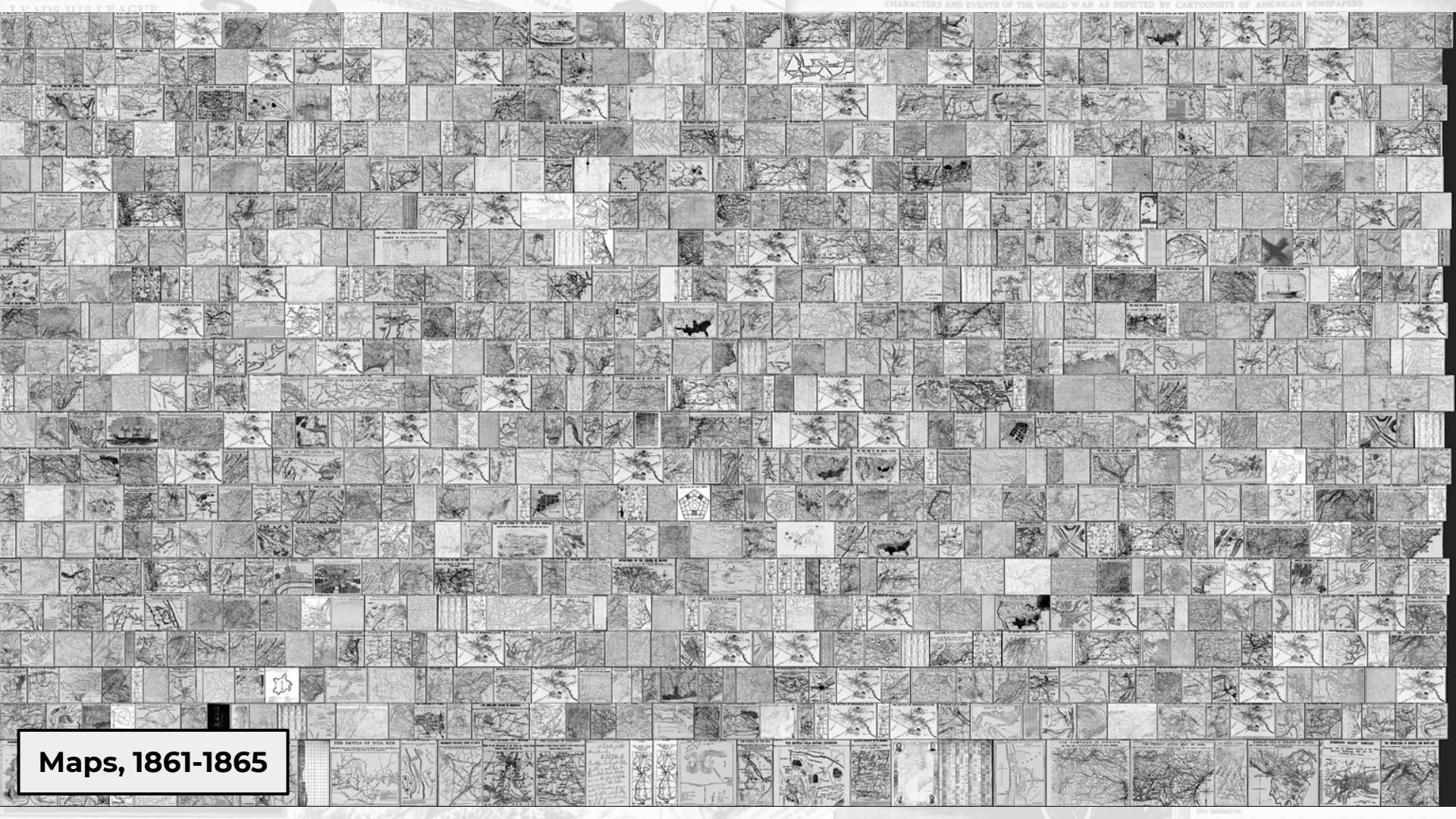
Dataset



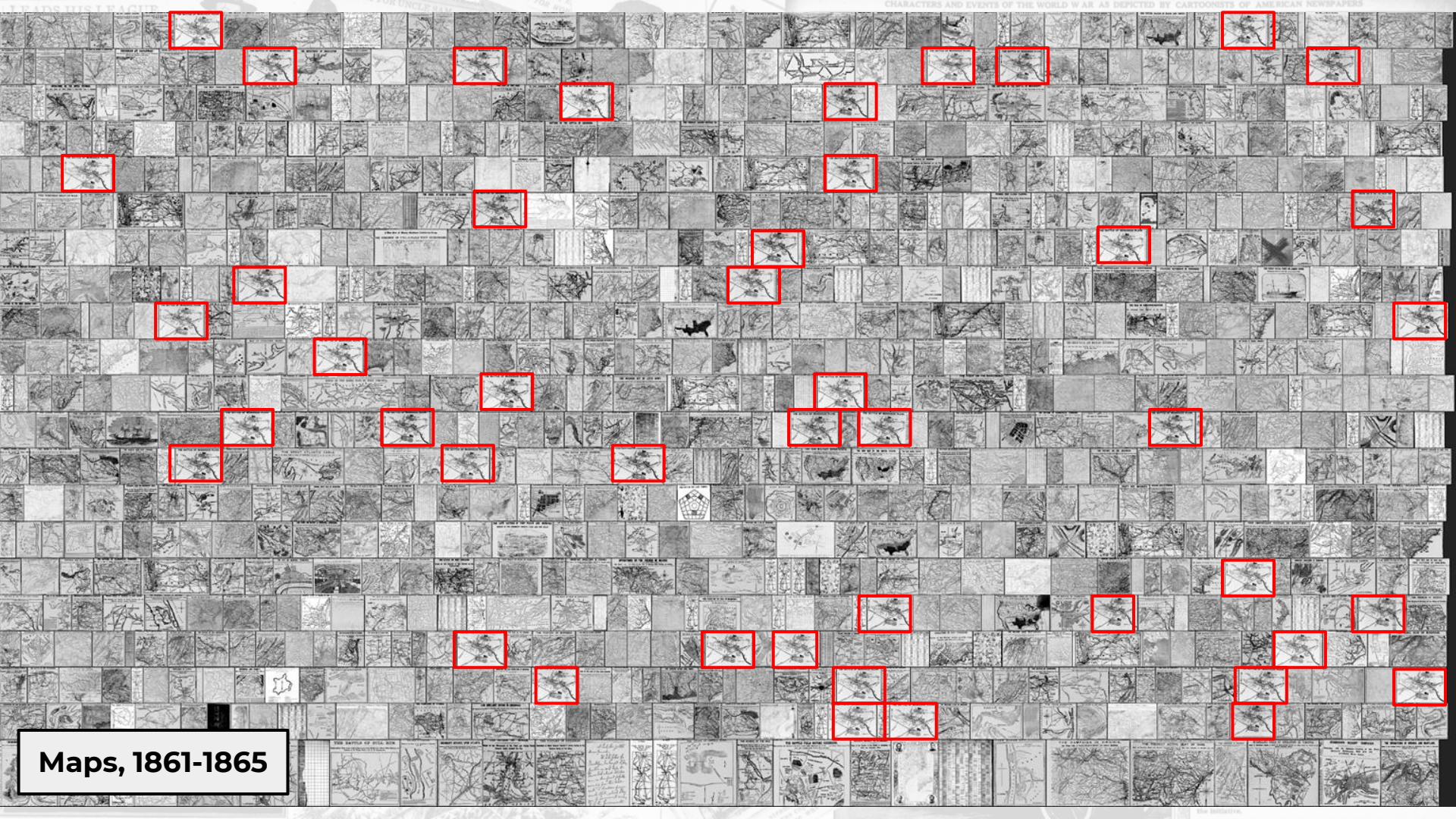
The Public!



Search Application



Maps, 1861-1865



Maps, 1861-1865

# Newspaper Navigator Search Application

Search My Collection Train My AI Navigators

Chronicling America LC Labs Data Archaeology About

Save Clear & Restart Gallery List

## Search with Newspaper Navigator!

Location:  Start Year:  End Year:

displaying 1 - 100 records in total 5427

Update Reset Filters

Sorting by:

Your AI Navigators  
None defined yet

baseball

displaying 1 - 100 records in total 5427

Update Reset Filters

Sorting by:

Your AI Navigators  
None defined yet

https://news-navigator.labs.loc.gov/search/select\_state=All&select\_start\_year=1900&select\_end\_year=1963&search=baseball&start\_time=167298112290&plus\_library=&positive=&negative=&view=gallery

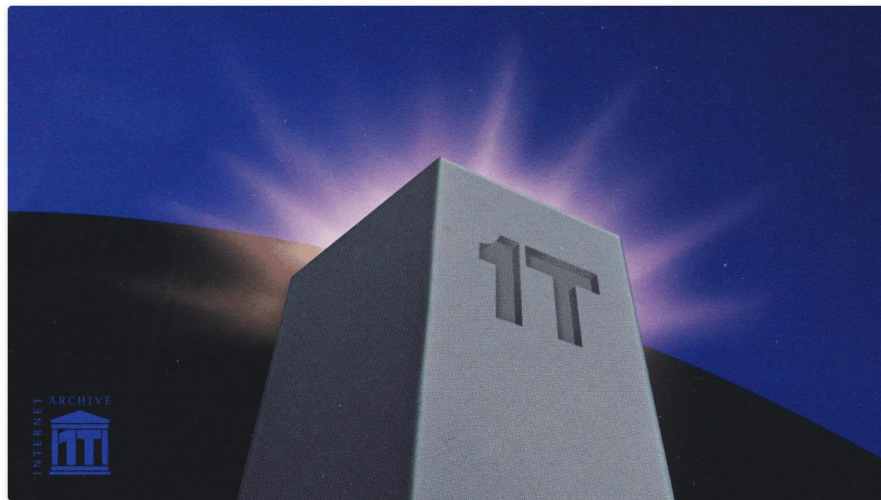
## About



# GovScape: Searching .gov Web Archives at Scale

# A New Frontier of Scale for Cultural Heritage

## Celebrating 1 Trillion Web Pages Archived



On October 22, 2025, the Internet Archive celebrated an extraordinary milestone: 1 trillion web pages preserved and available for access via the Wayback Machine.

<https://blog.archive.org/trillion/>

## Historians



# Historians

- “Our historical record is being and will be profoundly affected by the advent of these massive born-digital repositories.” - Ian Milligan, *History in the Age of Abundance*

# Historians

- “Our historical record is being and will be profoundly affected by the advent of these massive born-digital repositories.” - Ian Milligan, *History in the Age of Abundance*
- Particularly important as post-1990s topics enter the domain of historical research

# Historians

- “Our historical record is being and will be profoundly affected by the advent of these massive born-digital repositories.” - Ian Milligan, *History in the Age of Abundance*
- Particularly important as post-1990s topics enter the domain of historical research
- New questions, new subjects, and new methodologies

# Powell.pps (with Trevor Owens and Jonah Estess)

**IRAQ** FAILING TO DISARM

## Iraq Is Hiding Dangerous Weapons

After inspectors revealed his earlier biological program, Saddam built biological weapons factories on wheels

Instructions relayed to an officer in the field, January 2003

A: Capt. Ibrahim?  
B: I am with you, sir.  
A: Remove.. the expression... **nerve agents**... whenever it comes up.. in the wireless instructions.  
B: Consider it done, sir.



The diagram shows a mobile production facility for biological agents, consisting of several interconnected components: Active Material Tanks, Spray Dryers, Filling Machine, Fermentation, and Control Panel. The facility is mounted on a truck chassis.

U.S. SECRETARY OF STATE POWELL'S PRESENTATION TO THE UN SECURITY COUNCIL

## Powell.pps: Close and Distant Reading of Primary Sources in Web Archives

Trevor Owens, Benjamin Charles Germain Lee, Jonah Estess

Internet Histories, 2025: <https://doi.org/10.1080/24701475.2025.2516306>

# Powell.pps (with Trevor Owens and Jonah Estess)

- From “at scale” to the case study: a single artifact

## IRAQ FAILING TO DISARM

### Iraq Is Hiding Dangerous Weapons

After inspectors revealed his earlier biological program, Saddam built biological weapons factories on wheels



Instructions relayed to an officer in the field, January 2003

A: Capt. Ibrahim?

B: I am with you, sir.

A: Remove.. the expression... **nerve agents...** whenever it comes up.. in the wireless instructions.

B: Consider it done, sir.



U.S. SECRETARY OF STATE POWELL'S PRESENTATION TO THE UN SECURITY COUNCIL

# Powell.pps (with Trevor Owens and Jonah Estess)

- From “at scale” to the case study: a single artifact
- Tracing Colin Powell’s 2003 PowerPoint presentation to the United Nations Security Council

## IRAQ FAILING TO DISARM

### Iraq Is Hiding Dangerous Weapons

After inspectors revealed his earlier biological program, Saddam built biological weapons factories on wheels



Instructions relayed to an officer in the field, January 2003

A: Capt. Ibrahim?

B: I am with you, sir.

A: Remove.. the expression... **nerve agents...** whenever it comes up.. in the wireless instructions.

B: Consider it done, sir.



U.S. SECRETARY OF STATE POWELL'S PRESENTATION TO THE UN SECURITY COUNCIL

# Powell.pps (with Trevor Owens and Jonah Estess)

- From “at scale” to the case study: a single artifact
- Tracing Colin Powell’s 2003 PowerPoint presentation to the United Nations Security Council
- Demonstrating the value of searchability for born-digital government publications

**IRAQ** FAILING TO DISARM

## Iraq Is Hiding Dangerous Weapons

After inspectors revealed his earlier biological program, Saddam built biological weapons factories on wheels



Instructions relayed to an officer in the field, January 2003

A: Capt. Ibrahim?  
B: I am with you, sir.  
A: Remove.. the expression... **nerve agents...** whenever it comes up.. in the wireless instructions.  
B: Consider it done, sir.

U.S. SECRETARY OF STATE POWELL'S PRESENTATION TO THE UN SECURITY COUNCIL

# Legal Scholars

## PERMA: SCOPING AND ADDRESSING THE PROBLEM OF LINK AND REFERENCE ROT IN LEGAL CITATIONS

*Jonathan Zittrain, Kendra Albert, and Lawrence Lessig\**

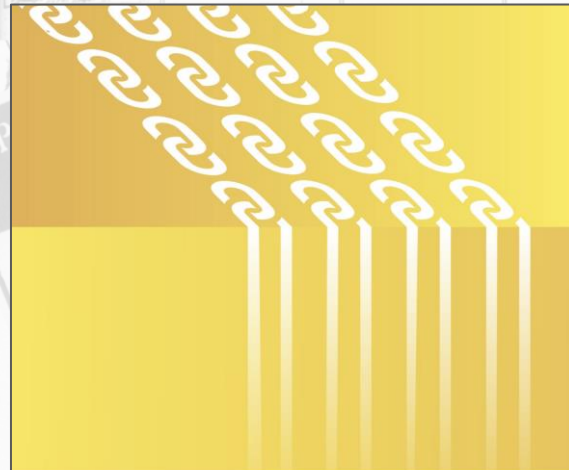
### INTRODUCTION

Works of scholarship have long cited primary sources or academic works to provide sources for facts, to incorporate previous scholarship, and to bolster arguments. The ideal citation connects an interested reader to what the author references, making it easy to track down, verify, and learn more from the indicated sources.

In principle, as cited sources move to the Web, this linking should become easier. Rather than requiring a reader to travel to a library to follow the sources cited by an author, the reader should be able to retrieve the cited material immediately with a single click.

But again, only in principle. The link, a URL, points to a resource hosted by a third party. That resource will only survive so as long as the third party preserves it. And as websites evolve, not all third parties will have a sufficient interest in preserving the links that provide backwards compatibility to those who relied upon those links. The author of the cited source may decide the argument in the source was mistaken and take it down. The website owner may decide to abandon one mode of organizing material for another. Or the organization providing the source material may change its views and “update” the original source to reflect its evolving views. In each case, the citing paper is vulnerable to footnotes that no longer support its claims. This vulnerability threatens the integrity of the resulting scholarship.

This problem does not exist for printed sources, or at least not in the same way. Print sources can be kept indefinitely by libraries or archives, assuming space and other determinations allow. The ability to update those original print sources is, for these purposes, happily difficult. Tracking down every original copy of an edition of a printed *New York Times* and changing a story on page A4 is the stuff of



## THE PAPER OF RECORD MEETS AN EPHEMERAL WEB: AN EXAMINATION OF LINKROT AND CONTENT DRIFT WITHIN THE *NEW YORK TIMES*

JOHN BOWERS • CLARE STANTON • JONATHAN ZITTRAIN

SPRING 2021

LIL.LAW.HARVARD.EDU



Electronic copy available at: <https://ssrn.com/abstract=3833133>

<https://harvardlawreview.org/forum/vol-127/perma-scoping-and-addressing-the-problem-of-link-and-reference-rot-in-legal-citations/>

[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3833133](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3833133)

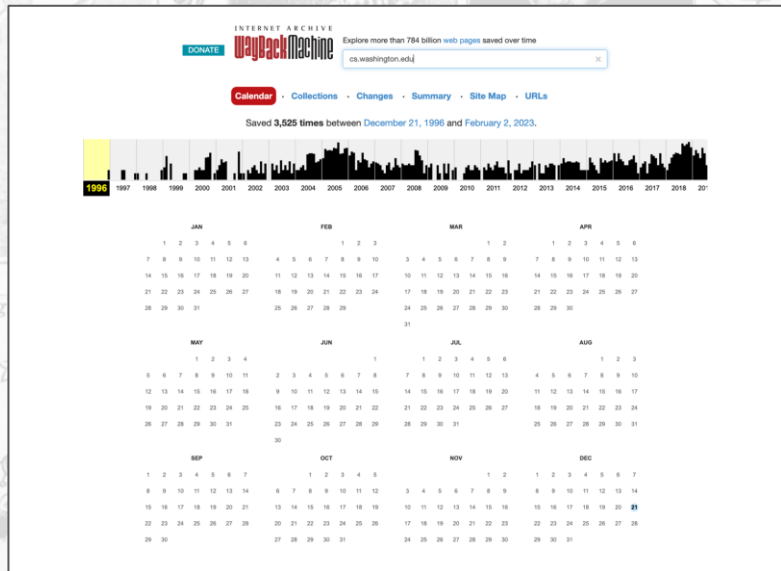
# Computer Scientists

- Already central to AI research!
- Training foundation models with web derivatives

| Dataset                 | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|-------------------------|-------------------|------------------------|----------------------------------------------|
| Common Crawl (filtered) | 410 billion       | 60%                    | 0.44                                         |
| WebText2                | 19 billion        | 22%                    | 2.9                                          |
| Books1                  | 12 billion        | 8%                     | 1.9                                          |
| Books2                  | 55 billion        | 8%                     | 0.43                                         |
| Wikipedia               | 3 billion         | 3%                     | 3.4                                          |

**Table 2.2: Datasets used to train GPT-3.** “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

# Web Archives



Wayback Machine http://www.cs.washington.edu/ 3,525 captures 1996-2023

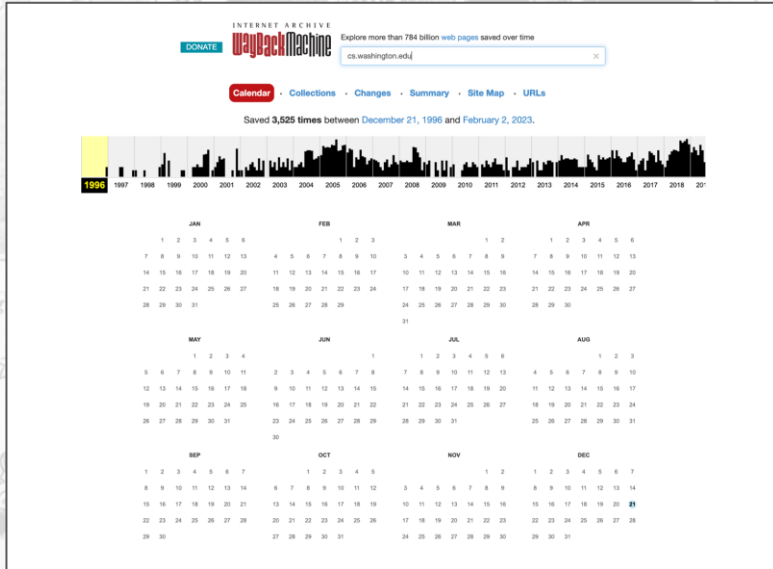
## University of Washington Computer Science & Engineering

Search Engine: Public Computer Science - 1996-2023

- GENERAL INFORMATION**  
Including an overview of the department, visitor schedule, admissions, internships, which are on our web, connection, research, and more.
- EDUCATION**  
Including a time schedule of classes, course list and syllabi, information about the full-time graduate program, the professional master program, and the undergraduate computer science and computer engineering programs, final exam schedules, and more.
- RESEARCH**  
Including research project web pages, technical reports and abstracts, Computing Research Association, and more.
- PEOPLE & ORGANIZATIONS**  
Including faculty, staff, students, visitors, organizations, our Affiliates Program, our graduate Ph.D. students, and more.
- THE REGION**  
Including local information, desktop references, links to documents, and more.
- SPOTLIGHT**
  - Professional Masters Program (Application deadline for Spring 1997, February 1)
  - UW wins Pacific Regionals of ACM International Student Programming Contest
  - Ten videos highlighting educational initiatives
  - Our collections are now live on the MROHE
  - Major donation from Intel Corporation
  - Dick Karp receives National Medal of Science
  - Department Overview
  - The Impact of a Research University: An IT Perspective
  - Faculty/Staff Positions Available
  - A Half Century of Exponential Progress in Information Technology? (1996 UW Annual Faculty Lecture), in transcripts or transcriptions plus RealAudio

# Web Archives

What might other modes of discoverability look like for web archives, and government information?



University of Washington Computer Science & Engineering website. The page includes a navigation menu with links to General Information, Education, Research, People & Organizations, The Region, and Spotlight. The main content area lists various programs and initiatives, including the Professional Masters Program, UW-wide Pacific Regional of ACM International Student Programming Contest, and the Dick Karp National Medal of Science.

Navigation menu: GENERAL INFORMATION, EDUCATION, RESEARCH, PEOPLE & ORGANIZATIONS, THE REGION, SPOTLIGHT

Main content area: Including an overview of the department, visitor schedule, graduate information, and more. Including a time schedule of classes, course list and syllabi, information about the full-time graduate program, the professional masters program, and the undergraduate computer science and computer engineering programs, final exam schedules, and more. Including research project web pages, technical reports and abstracts, Computing Research Association, and more. Including faculty, staff, students, visitors, organizations, our Affiliates Program, our graduating Ph.D. students, and more. Including local information, desktop references, links to documents, and more. Professional Masters Program (Application deadline for Spring 1997: February 1). UW-wide Pacific Regional of ACM International Student Programming Contest. Two videos highlighting educational initiatives. Our collections are now live on the Mbone. Major donation from Intel Corporation. Dick Karp receives National Medal of Science. Department Overview. The Impact of a Research University: An IT Perspective. Faculty/Staff Positions Available. A Half Century of Exponential Progress in Information Technology? (1996 UW Annual Faculty Lecture), in transcriptions or transcriptions plus RealAudio.

# Initial Findings

## Grappling with the Scale of Born-Digital Government Publications: Toward Pipelines for Processing and Searching Millions of PDFs

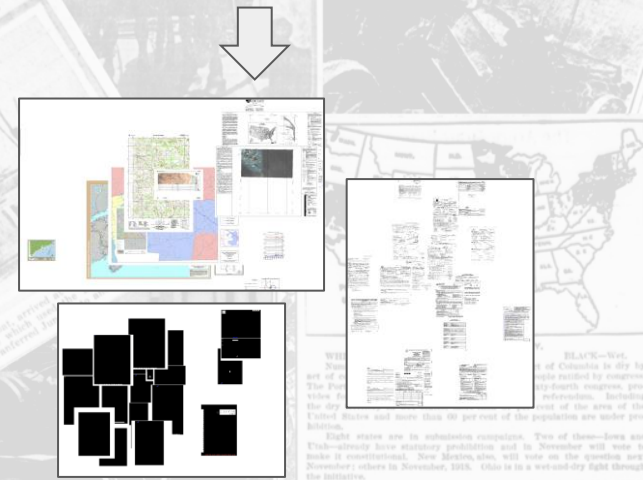
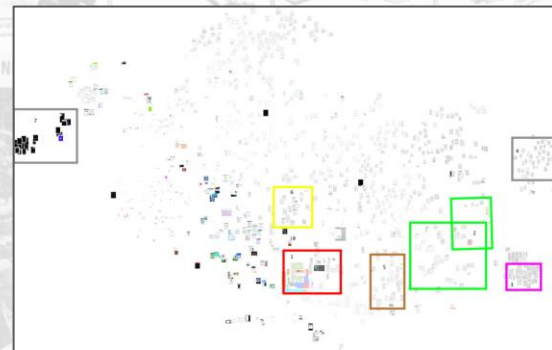
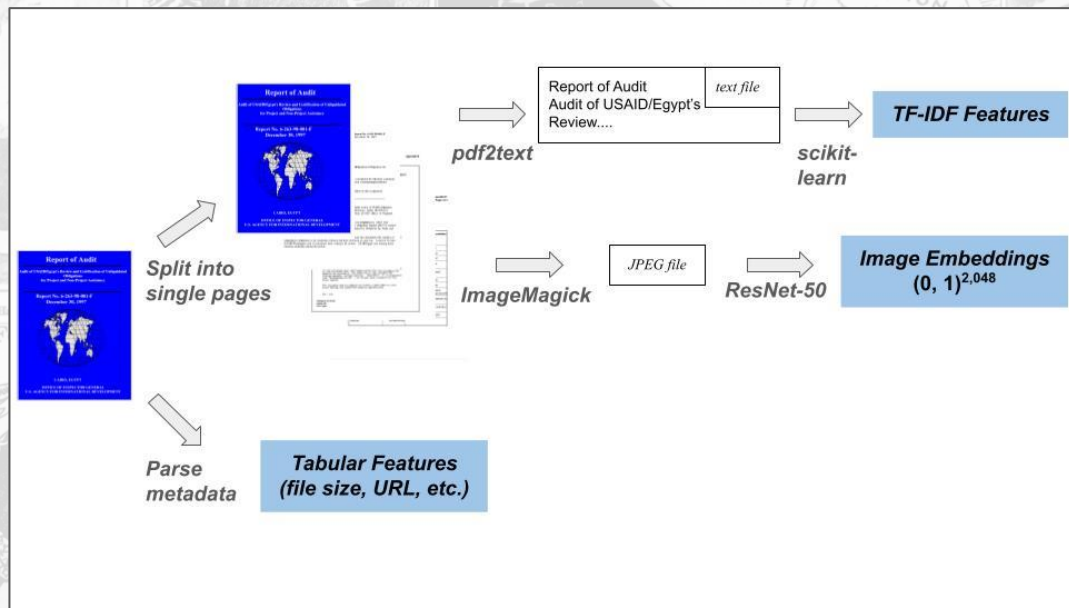
Benjamin Charles Germain Lee and Trevor Owens

*International Journal of Digital Humanities*, Volume 3 (2022)



<https://arxiv.org/abs/2112.02471>  
<https://doi.org/10.1007/s42803-022-00042-x>

# A Scalable Multimodal Pipeline



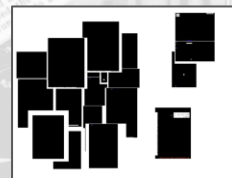
# A Scalable Multimodal Pipeline

Where can we  
expand this?

Split into  
single pages

Parse  
metadata

Tabular Features  
(file size, URL, etc.)



# End of Term Web Archive

End of Term Web Archive

Background Partners Datasets

## Purpose

The End of Term Web Archive captures and saves U.S. Government websites at the end of presidential administrations. The EOT has thus far preserved websites from administration changes in 2008, 2012, 2016, and 2020.



Whitehouse.gov captures from: September 15, 2008; March 21, 2013; and February 3, 2017.

## Archive Scope

The End of Term Web Archive contains federal government websites (.gov, .mil, etc) in the Legislative, Executive, or Judicial branches of the government. Websites that were at risk of changing (i.e., whitehouse.gov) or disappearing altogether during government transitions were captured. Local government websites, or any other site not part of the federal government domain were out of scope.

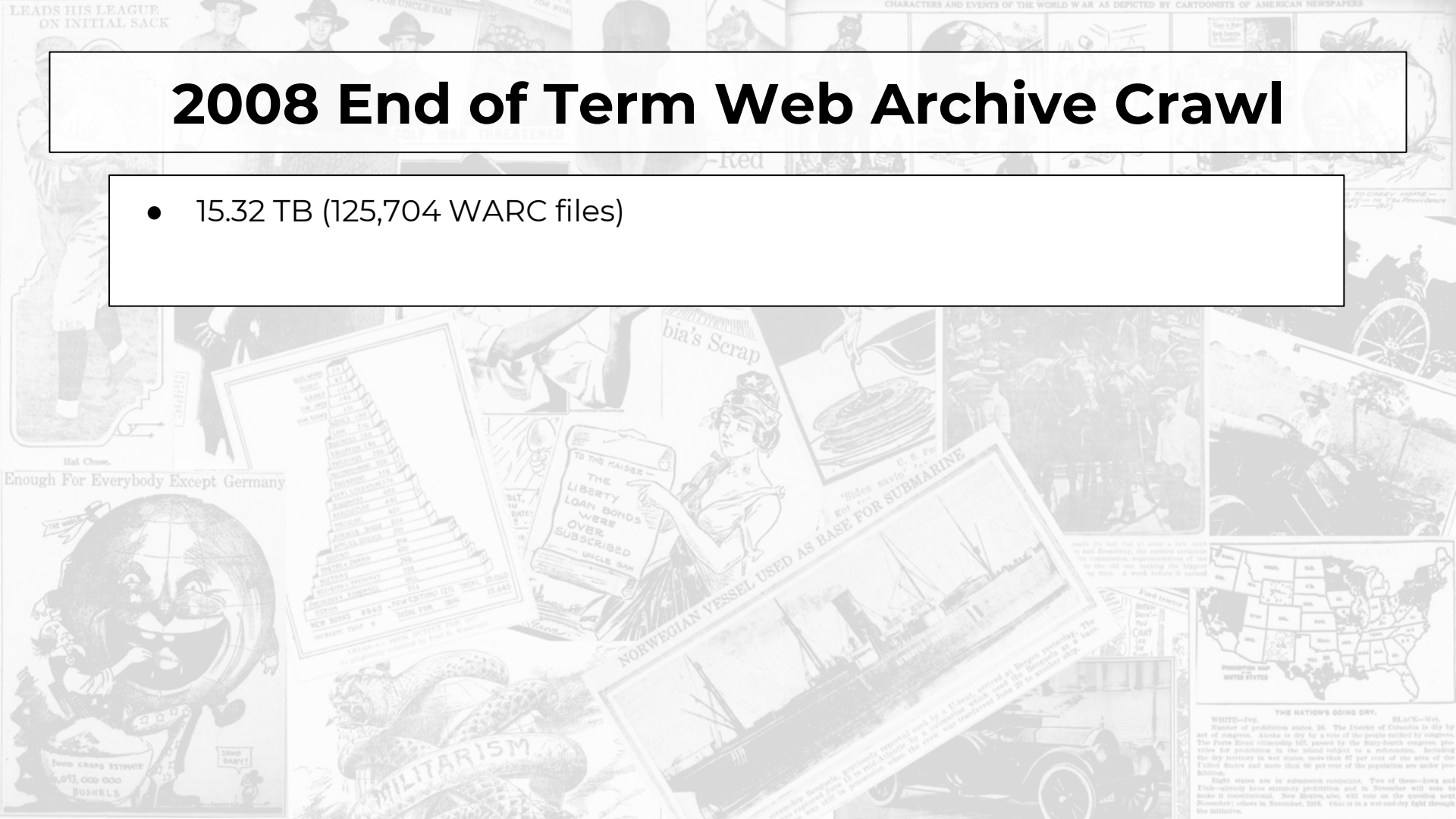
<https://eotarchive.org/>

# 2008 End of Term Web Archive Crawl



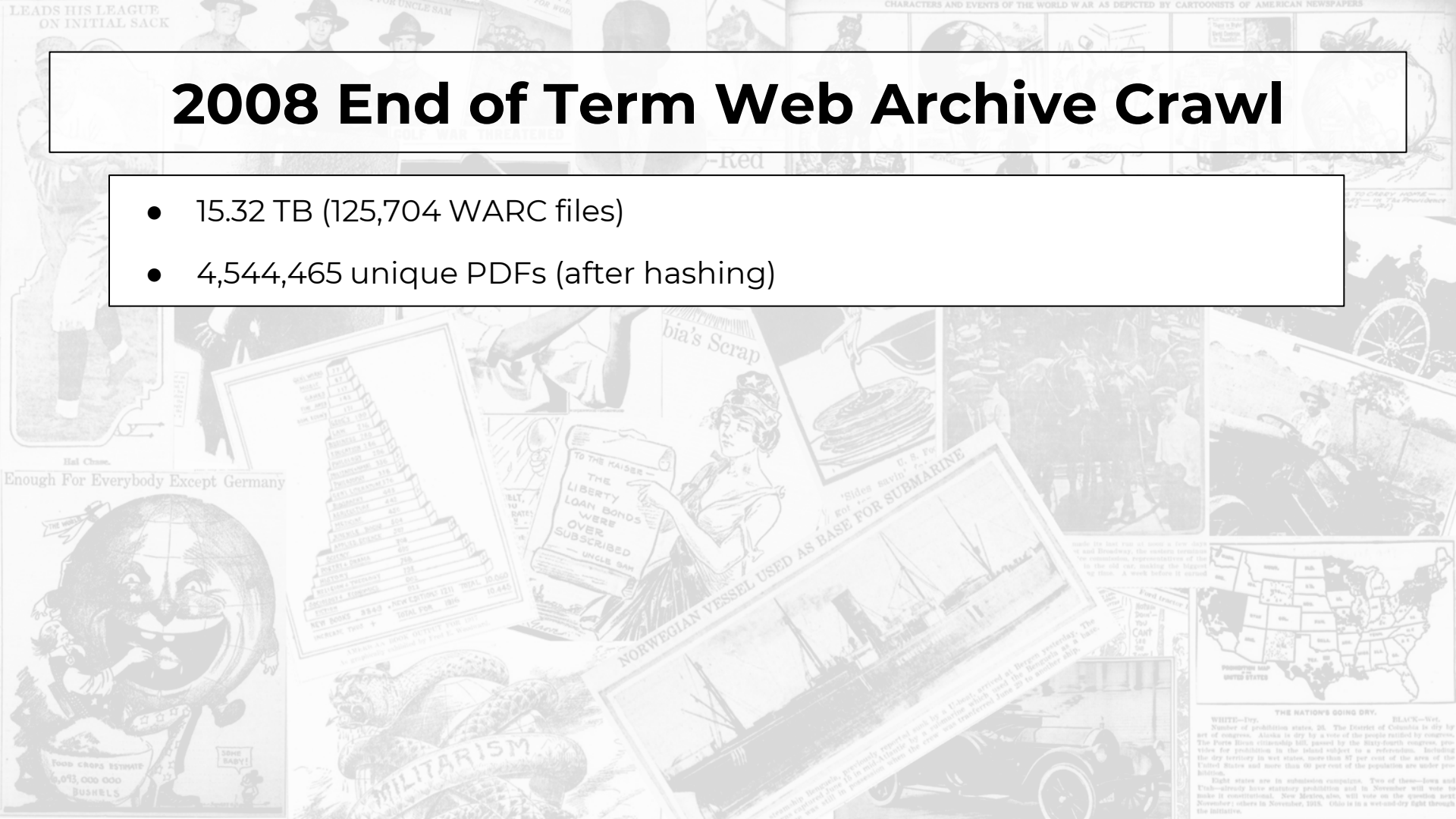
# 2008 End of Term Web Archive Crawl

- 15.32 TB (125,704 WARC files)



# 2008 End of Term Web Archive Crawl

- 15.32 TB (125,704 WARC files)
- 4,544,465 unique PDFs (after hashing)



# 2008 End of Term Web Archive Crawl

- 15.32 TB (125,704 WARC files)
- 4,544,465 unique PDFs (after hashing)

Table 6. Top Level Subdomains by PDF Documents and Pages

| Rank | Total # Documents | # One-page Documents | # Documents >= 20 Pages |
|------|-------------------|----------------------|-------------------------|
| 1    | gpo.gov           | gpo.gov              | gpo.gov                 |
| 2    | usda.gov          | usda.gov             | gao.gov                 |
| 3    | house.gov         | house.gov            | epa.gov                 |
| 4    | army.mil          | uscis.gov            | usda.gov                |
| 5    | bea.gov           | uscourts.gov         | army.mil                |
| 6    | census.gov        | army.mil             | noaa.gov                |

Table 7. Distribution of PDF Documents by Range of Pages

| Page Range | #         | %      | Cumulative % |
|------------|-----------|--------|--------------|
| 1          | 1,477,612 | 33.55% | 33.55%       |
| 2-14       | 2,203,216 | 50.03% | 83.58%       |
| 15-100     | 616,552   | 14.00% | 97.58%       |
| 101-1,000  | 104,766   | 2.38%  | 99.96%       |
| 1,001+     | 1,902     | 0.04%  | 100.00%      |

Phillips, Mark Edward & Murray, Kathleen R. Improving Access to Web Archives through Innovative Analysis of PDF Content, paper, April 2013;  
(<https://digital.library.unt.edu/ark:/67531/metadc155622/>, University of North Texas Libraries)

# 2020 End of Term Web Archive Crawl

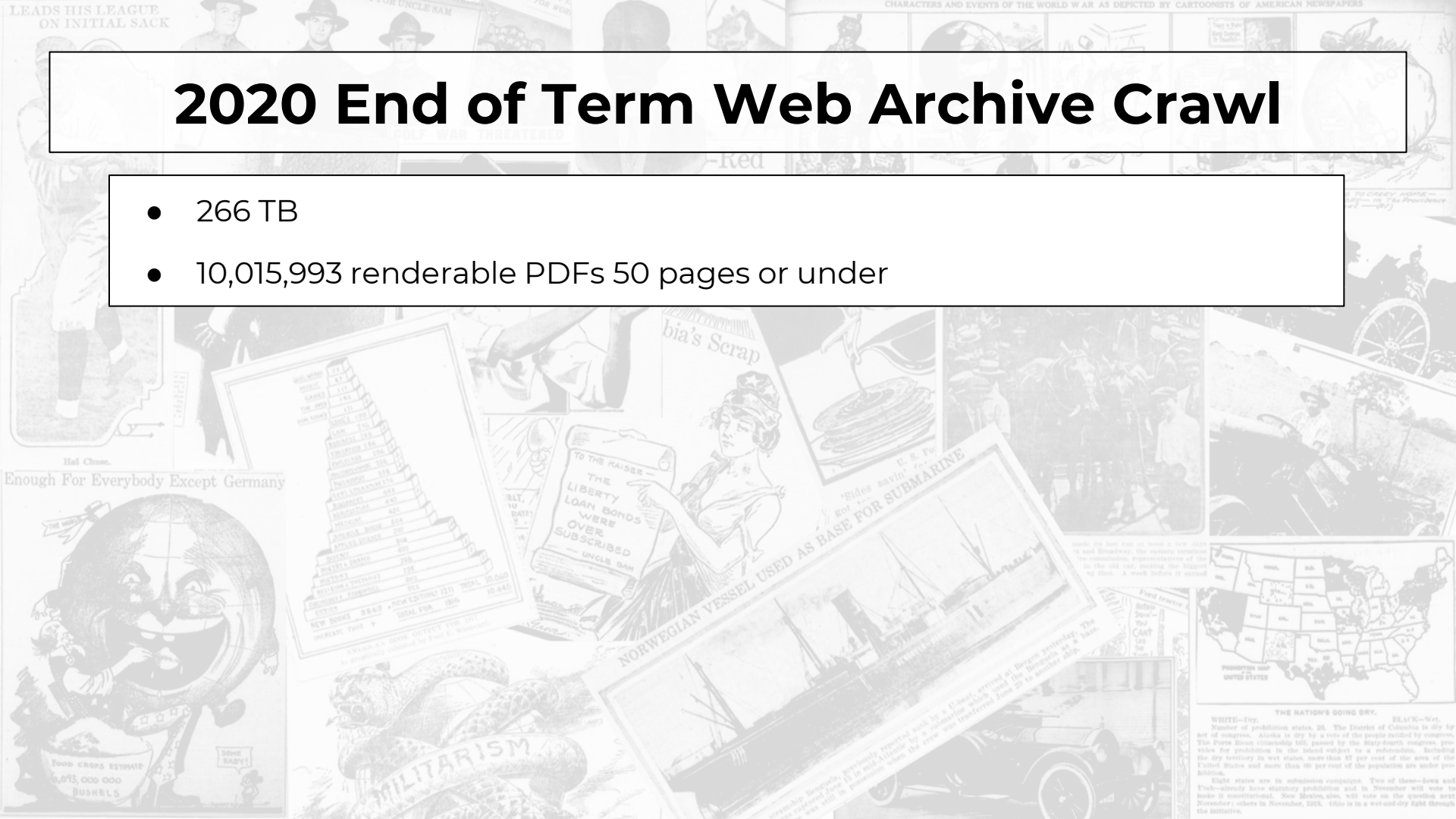
2020 End of Term Web Archive Crawl

- 266 TB



# 2020 End of Term Web Archive Crawl

- 266 TB
- 10,015,993 renderable PDFs 50 pages or under



# 2020 End of Term Web Archive Crawl

- 266 TB
- 10,015,993 renderable PDFs 50 pages or under

| Domain      | # of PDF pages | # of PDFs |
|-------------|----------------|-----------|
| sec.gov     | 7,324,033      | 791,197   |
| govinfo.gov | 3,051,220      | 1,078,761 |
| ca.gov      | 2,860,263      | 375,405   |
| bls.gov     | 2,261,400      | 234,667   |
| wa.gov      | 2,210,997      | 248,777   |
| usda.gov    | 2,149,605      | 311,660   |
| epa.gov     | 1,990,902      | 186,806   |
| ed.gov      | 1,978,199      | 157,778   |
| nasa.gov    | 1,954,285      | 192,714   |

|                     |           |         |
|---------------------|-----------|---------|
| noaa.gov            | 1,770,684 | 195,122 |
| cdc.gov             | 1,493,247 | 157,830 |
| illinois.gov        | 1,212,092 | 121,004 |
| uscourts.gov        | 1,178,300 | 141,622 |
| federalregister.gov | 1,128,733 | 162,980 |
| fcc.gov             | 1,113,856 | 183,620 |
| uspto.gov           | 1,026,527 | 104,719 |
| utah.gov            | 962,937   | 200,700 |
| nih.gov             | 959,782   | 97,499  |
| census.gov          | 914,647   | 223,834 |
| hawaii.gov          | 894,776   | 157,264 |

TABLE I: A breakdown of the 20 most prevalent domains in GovScope, along with the associated number of PDFs included. We note that counts in this table (and this table alone) reflect PDFs before de-duplication.

## GovScape: Goals



# GovScape: Goals

- Developing **multimodal search** for these government PDFs



# GovScape: Goals

- Developing **multimodal search** for these government PDFs
- Serving government PDFs **at scale**



# GovScape: Goals

- Developing **multimodal search** for these government PDFs
- Serving government PDFs **at scale**
- Understanding how **patrons**, from journalists to scholars to policy analysts, want to search gov docs with this system



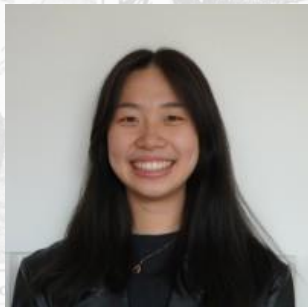
# GovScape: Goals

- Developing **multimodal search** for these government PDFs
- Serving government PDFs **at scale**
- Understanding how **patrons**, from journalists to scholars to policy analysts, want to search gov docs with this system
- Ongoing: scaling out to cover the entire End of Term Web Archive



# GovScape

And  
others!



Alison Yan



Claire Gong



Kyle Deeds



Trevor Owens



Mark Phillips



Shannon Shen



SJ Klein



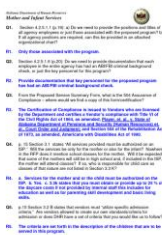
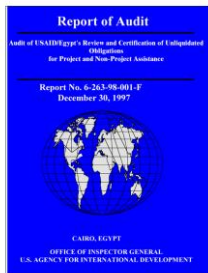
Ying-Hsiang Huang



Leslie Harka



Akshay Mehta



# GovScape

## Semantic Text Search

(“pediatric healthcare in rural areas”)

## Visual Search

(“redacted documents”)

## Keyword Search

(“Nancy Pelosi”)

## Filter Conditions / Metadata Facets

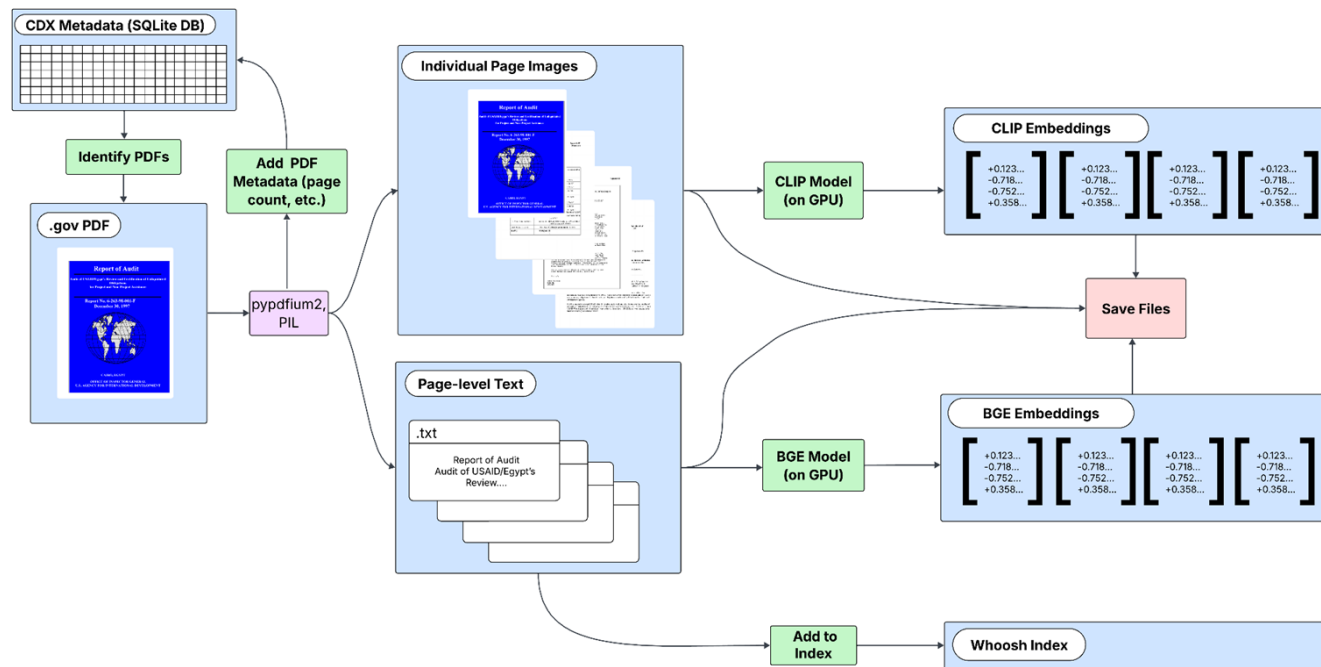
(URL, crawl date)

~10 million PDFs,  
~100 million PDF pages

## Two Key Constraints:

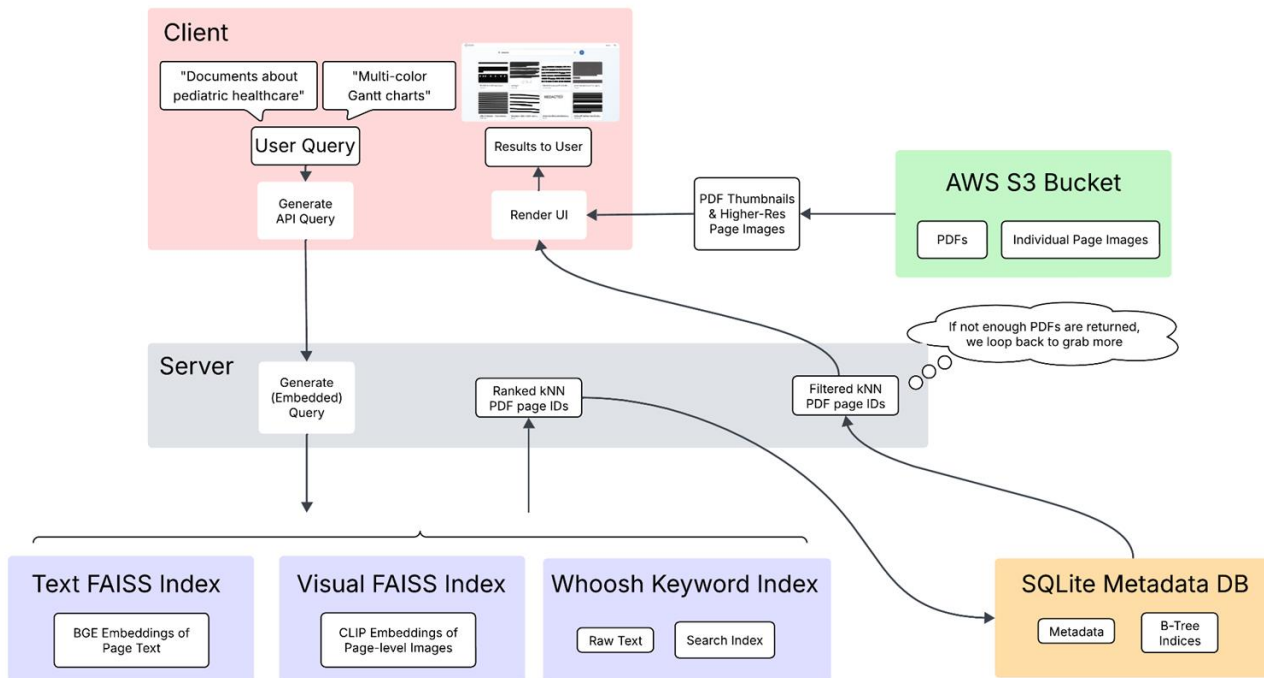
- Low latency
- Computing infrastructure on an academic budget

# GovScape



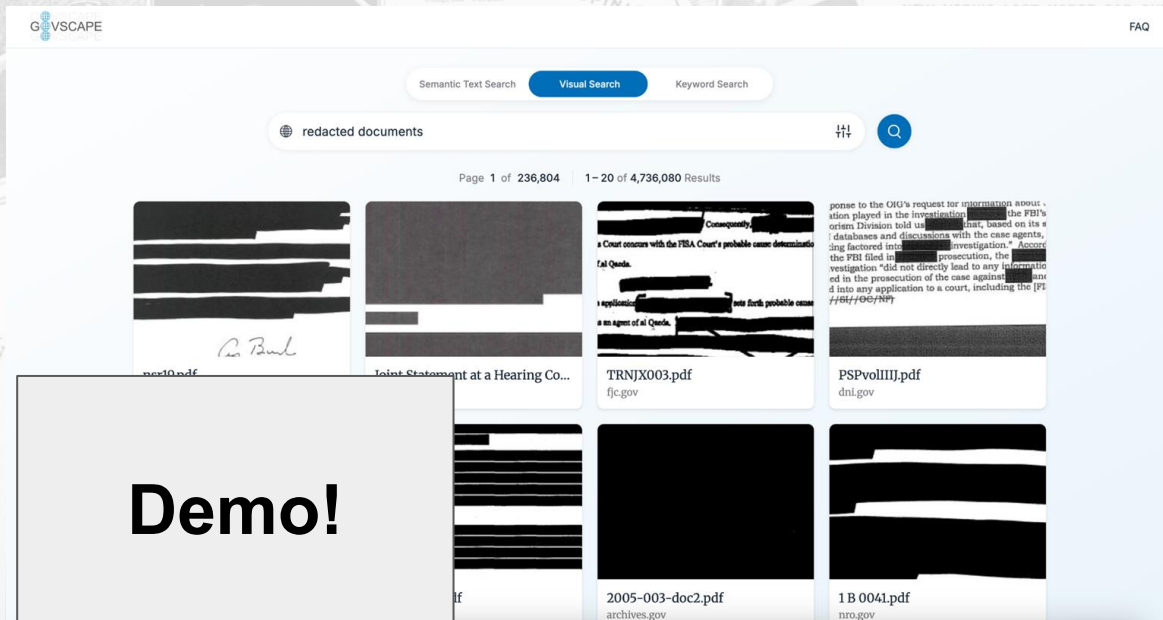
<https://www.govscape.net/>

# GovScape



<https://www.govscope.net/>

# GovScape



<https://www.govscope.net/>



| Task Type                                                            | EC2 Instance Type | Time           | Cost              |
|----------------------------------------------------------------------|-------------------|----------------|-------------------|
| Pipeline Minus Indexing<br>(prior run with 4,736,080 PDFs)           | g4dn.4xlarge      | 515.67 hours   | \$620.87          |
| 1. List                                                              |                   | 8.62 hours     |                   |
| 2. Download                                                          |                   | 19.31 hours    |                   |
| 3. PDF to Text and Image                                             |                   | 89.16 hours    |                   |
| 4. Text Embedding                                                    |                   | 274.49 hours   |                   |
| 5. Image Embedding                                                   |                   | 100.00 hours   |                   |
| 6. Metadata                                                          |                   | 1.95 hours     |                   |
| 7. Upload                                                            |                   | 22.14 hours    |                   |
| <b>Pipeline Minus Indexing<br/>(Extrapolated to 10,015,993 PDFs)</b> | g4dn.4xlarge      | 1,090.55 hours | <b>\$1,313.02</b> |
| <b>FAISS Index Building<br/>for 10,015,993 PDFs</b>                  | g4dn.4xlarge      | 54.64 hours    | <b>\$65.79</b>    |
| <b>Keyword Index Building<br/>for 10,015,993 PDFs</b>                | m5.4xlarge        | 38.79 hours    | <b>\$29.79</b>    |
| <b>Total pre-processing (Extrapolated)</b>                           |                   |                | <b>\$1,408.60</b> |

TABLE II: A breakdown of compute for the GovScape pre-processing pipeline, including task type, compute specifications, time, and cost. All compute is reported using AWS g4dn.4xlarge instances (each with 16 Intel Cascade Lake vCPUs & 1 Nvidia T4 GPU) and m5.4xlarge instances (each with 16 Intel Xeon Platinum 8175 vCPUs) in the us-east-2 region. We have separated out index building in this table because each index can only be built by a single server at a time. For the other steps, we extrapolate out our compute cost using a previous run of 4,736,080 PDFs, which we had configured for benchmarking.

| Task Type                                                            | EC2 Instance Type | Time         | Cost              |
|----------------------------------------------------------------------|-------------------|--------------|-------------------|
| Pipeline Minus Indexing<br>(prior run with 4,736,080 PDFs)           | g4dn.4xlarge      | 515.67 hours | \$620.87          |
| 1. List                                                              |                   | 8.62 hours   |                   |
| 2. Download                                                          |                   | 19.31 hours  |                   |
| 3. PDF to Text and Image                                             |                   | 80.16 hours  |                   |
| 4. Text Embedding                                                    |                   |              |                   |
| 5. Image Embedding                                                   |                   |              |                   |
| 6. Metadata                                                          |                   |              |                   |
| 7. Upload                                                            |                   |              |                   |
| <b>Pipeline Minus Indexing<br/>(Extrapolated to 10,015,993 PDFs)</b> | g4dn.4xlarge      |              |                   |
| <b>FAISS Index Building<br/>for 10,015,993 PDFs</b>                  | g4dn.4xlarge      |              |                   |
| <b>Keyword Index Building<br/>for 10,015,993 PDFs</b>                | m5.4xlarge        | 38.79 hours  | <b>\$29.79</b>    |
| <b>Total pre-processing (Extrapolated)</b>                           |                   |              | <b>\$1,408.60</b> |

- **47,000+ PDF pages per \$1 spent on compute**
- **Scalable to a billion pages**

TABLE II: A breakdown of compute for the GovScope pre-processing pipeline, including task type, compute specifications, time, and cost. All compute is reported using AWS g4dn.4xlarge instances (each with 16 Intel Cascade Lake vCPUs & 1 Nvidia T4 GPU) and m5.4xlarge instances (each with 16 Intel Xeon Platinum 8175 vCPUs) in the us-east-2 region. We have separated out index building in this table because each index can only be built by a single server at a time. For the other steps, we extrapolate out our compute cost using a previous run of 4,736,080 PDFs, which we had configured for benchmarking.

# Paper at Association for Computational Linguistics (ACL) 2026

## GovScape: A Public Multimodal Search System for 70 Million Pages of Government PDFs

Ying-Hsiang Huang, Claire Gong, Shreya Shaji, Alison R Yan, Leslie Harka, Albert Du, Anjali Shubha Gopal, Samuel J Klein, Shannon Zejiang Shen, Mark E. Phillips, Trevor Owens, Kyle Deeds, Benjamin Charles Germain Lee

### Abstract

Efforts over the past three decades have produced web archives containing billions of webpage snapshots and petabytes of data. The End of Term Web Archive alone contains millions of PDFs produced by the federal government. While preservation with web archives has been successful, significant challenges for access and discoverability remain. In this paper, we introduce GovScape, a public search system that supports multimodal searches across 10,015,993 federal government PDFs from the 2020 End of Term crawl (70,958,487 total PDF pages) – to our knowledge, all renderable PDFs in the 2020 crawl that are 50 pages or under. GovScape supports four primary forms of search: in addition to providing (1) filter conditions over metadata facets including domain and crawl date and (2) exact text search against the PDF text, we provide (3) semantic text search and (4) visual search against the PDFs across individual pages, enabling users to structure queries such as “redacted documents” or “pie charts.” We detail GovScape’s search affordances, embedding pipeline, system architecture, and open source codebase. Significantly, the total estimated compute cost for GovScape’s pre-processing pipeline for 10 million PDFs was approximately 1,500, equivalent to 47,000 PDF pages per dollar spent on compute, demonstrating the potential for immediate scalability. We evaluate GovScape by (1) analyzing 1,679 search queries and (2) benchmarking vector and keyword index efficiency using these queries. GovScape can be found at <https://www.govscope.net>.

PDF

Cite

Search

Fix data



**Paper:** <https://aclanthology.org/2026.acl-demo.23/>  
**Codebase:** <https://github.com/govscope/govscope>

[illegible]

# Usage Statistics

- 3,000+ queries in first ~six months

# Usage Statistics

- **3,000+ queries** in first ~six months
- **Textual:** “Property tax survey,” “contaminant hazard reviews,” “policies regarding rainwater collection”

# Usage Statistics

- **3,000+ queries** in first ~six months
- **Textual:** “Property tax survey,” “contaminant hazard reviews,” “policies regarding rainwater collection”
- **Visual:** “Panoramic,” “network,” “data tables,” “maps of Genoa,” “Geologic map of the Chisos Mountain, Big Bend National Park”

# Usage Statistics

- **3,000+ queries** in first ~six months
- **Textual:** “Property tax survey,” “contaminant hazard reviews,” “policies regarding rainwater collection”
- **Visual:** “Panoramic,” “network,” “data tables,” “maps of Genoa,” “Geologic map of the Chisos Mountain, Big Bend National Park”
- **Keyword:** “NIST,” “artificial intelligence”

# GovText: A Dataset of 142 Billion Tokens

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

# GovText: A Dataset of 142 Billion Tokens

- OCR from 9+ million PDFs from the 2020 End of Term crawl

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

# GovText: A Dataset of 142 Billion Tokens

- OCR from 9+ million PDFs from the 2020 End of Term crawl
- 142.4 billion tokens of text (+ benchmarking against embedded text)

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

# GovText: A Dataset of 142 Billion Tokens

- OCR from 9+ million PDFs from the 2020 End of Term crawl
- 142.4 billion tokens of text (+ benchmarking against embedded text)
- All computationally accessible via Source Cooperative & AWS S3

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

# GovText: A Dataset of 142 Billion Tokens

- OCR from 9+ million PDFs from the 2020 End of Term crawl
- 142.4 billion tokens of text (+ benchmarking against embedded text)
- All computationally accessible via Source Cooperative & AWS S3
- Dataset available at: <https://source.coop/govscope/eota-ocr>

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

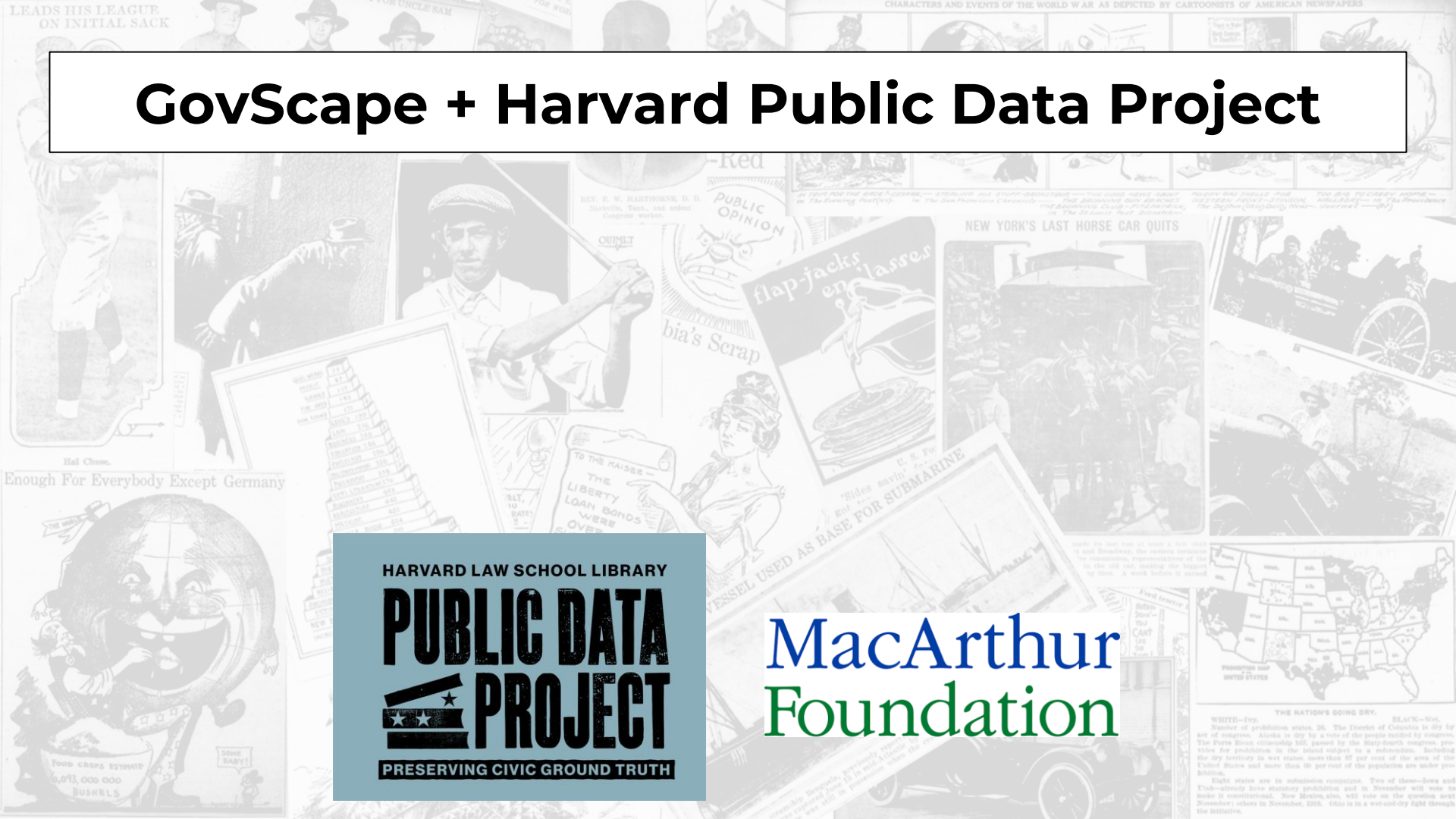
# GovText: A Dataset of 142 Billion Tokens

- OCR from 9+ million PDFs from the 2020 End of Term crawl
- 142.4 billion tokens of text (+ benchmarking against embedded text)
- All computationally accessible via Source Cooperative & AWS S3
- Dataset available at: <https://source.coop/govscope/eota-ocr>
- Paper forthcoming this month!

TABLE I: Summary statistics of GovText, with token counts computed for each text extraction method.

|                                          |               |
|------------------------------------------|---------------|
| Documents (PDFs)                         | 9,020,932     |
| Pages                                    | 169,989,592   |
| Pages per document, median (mean)        | 3 (18.84)     |
| Hosting URLs                             | 10,793,141    |
| Hosting URLs per document, median (mean) | 1 (1.20)      |
| olmOCR tokens                            | 142.4 Billion |
| pypdfium2 tokens                         | 91.1 Billion  |

# GovScape + Harvard Public Data Project



HARVARD LAW SCHOOL LIBRARY

**PUBLIC DATA  
PROJECT**

**PRESERVING CIVIC GROUND TRUTH**

MacArthur  
Foundation

# GovScape + Harvard Public Data Project

- Partnering with the Harvard Library Innovation Lab + Public Data Project

HARVARD LAW SCHOOL LIBRARY

**PUBLIC DATA  
PROJECT**

**PRESERVING CIVIC GROUND TRUTH**

MacArthur  
Foundation

# GovScape + Harvard Public Data Project

- Partnering with the Harvard Library Innovation Lab + Public Data Project
- Expanding multimodal search & discovery for new public data types, including ~700 TB of Smithsonian data

HARVARD LAW SCHOOL LIBRARY

**PUBLIC DATA  
PROJECT**

**PRESERVING CIVIC GROUND TRUTH**

MacArthur  
Foundation

# GovScape + Harvard Public Data Project

- Partnering with the Harvard Library Innovation Lab + Public Data Project
- Expanding multimodal search & discovery for new public data types, including ~700 TB of Smithsonian data
- Developing interoperable tooling (Digital Collections Explorer) to support this discoverability for heterogeneous collections

HARVARD LAW SCHOOL LIBRARY

**PUBLIC DATA  
PROJECT**

**PRESERVING CIVIC GROUND TRUTH**

MacArthur  
Foundation

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents



# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- Expanding our multimodal capabilities:

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data
  - Incorporating VLMs & RAG-style search

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data
  - Incorporating VLMs & RAG-style search
- **Developing interoperable tooling for multimodal entity tracking**

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data
  - Incorporating VLMs & RAG-style search
- **Developing interoperable tooling for multimodal entity tracking**
- **Exploring two case studies in the humanities:**

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data
  - Incorporating VLMs & RAG-style search
- **Developing interoperable tooling for multimodal entity tracking**
- **Exploring two case studies in the humanities:**
  - Literary studies: tracing the evolving portrayals of prominent American writers to understand their reception, legacies, and impact

# Toward a Multimodal Understanding of 250 Years of U.S. Government Documents

- **Expanding our multimodal capabilities:**
  - Developing a dataset of tens of millions of image-caption pairs
  - Finetuning CLIP + other models on gov data
  - Incorporating VLMs & RAG-style search
- **Developing interoperable tooling for multimodal entity tracking**
- **Exploring two case studies in the humanities:**
  - Literary studies: tracing the evolving portrayals of prominent American writers to understand their reception, legacies, and impact
  - History of science: studying the circulation of grey literature to understand evolving governmental narratives around scientific discovery and innovation

[illegible]

# GovScape: Long-term Goals

- A series of **user evaluations** surrounding our search affordances.

# GovScape: Long-term Goals

- A series of **user evaluations** surrounding our search affordances.
- **Collaborations** with **journalists**, **librarians**, **social scientists**, and **humanists** surrounding the use of this project to advance downstream questions from different disciplines and professions.

# GovScape: Long-term Goals

- A series of **user evaluations** surrounding our search affordances.
- **Collaborations** with **journalists**, **librarians**, **social scientists**, and **humanists** surrounding the use of this project to advance downstream questions from different disciplines and professions.
- Incorporating **mixed file types** (webpages, spreadsheets, etc.) from the End of Term Web Archive.

# GovScape: Long-term Goals

- A series of **user evaluations** surrounding our search affordances
- **Collaborations** with **journalists**, **librarians**, **social scientists**, and **humanists** surrounding the use of this project to advance downstream questions from different disciplines and professions
- Incorporating **mixed file types** (webpages, spreadsheets, etc.) from the End of Term Web Archive
- Developing **petabyte-scale search infrastructure**

# Speaking with you!

- We would love to learn from your expertise surrounding government information!
- We welcome any feedback and questions, and if you would be interested in speaking more, just let us know!
- Email: [bcgl@uw.edu](mailto:bcgl@uw.edu)



Information School  
UNIVERSITY of WASHINGTON

# Questions?

[illegible]

-Red