



INTERNET ARCHIVE

# The Annual Wikimania Update

Mark Graham  
September 18th, 2023



# Team “Turn All References Blue” (TARB)



Stephen Balbach



Mark Graham



Max Doerr



Jake Orlowitz



James Hare



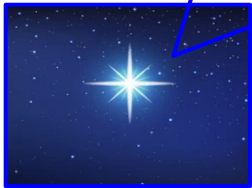
Dr. Sawood Alam



# Turn All References Blue



In the future nearly everything ever written or recorded, and all related resources, will be a click, gesture or a thought away.



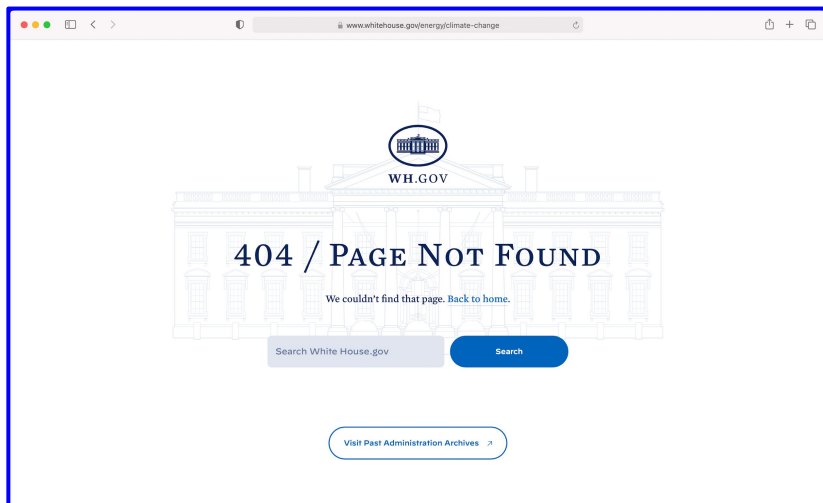
**The goal of the “Turn All References Blue” project is to contribute to making that a reality, starting with Wikipedia.**



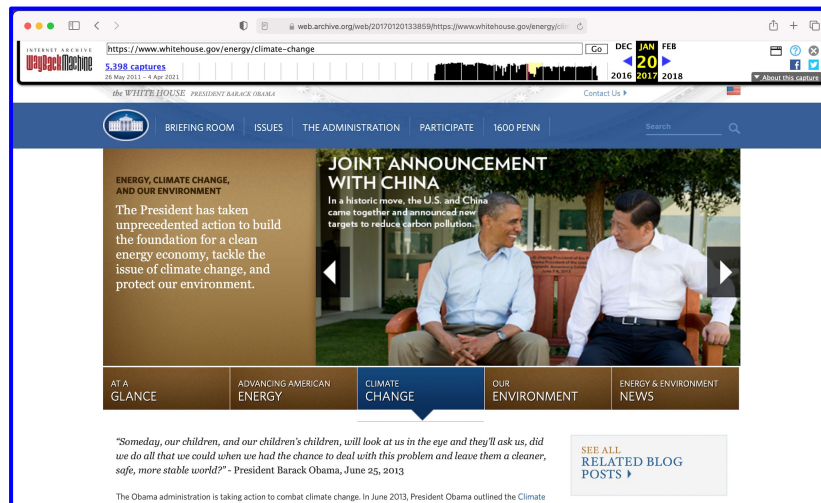
# Link Rot: When Good URLs go Bad

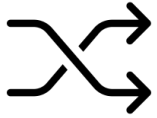


## Live Web (404)



## Archived Version





# Content Drift: Change Happens



April 2, 2020

April 6, 2020

Strategic National Stockpile

SNS Responds to the COVID-19 Pandemic

Strategic National Stockpile is the nation's largest supply of life-saving pharmaceuticals and medical supplies for use in a public health emergency severe enough to cause local supplies to run out.

When state, local, tribal, and territorial responders request federal assistance to support their response efforts, the stockpile ensures that the right medicines and supplies get to those who need them most during an emergency. Organized for scalable response to a variety of public health threats, this repository contains enough supplies to respond to multiple large-scale emergencies simultaneously.

About the Stockpile | Sustaining the Stockpile | Stockpile Products

Strategic National Stockpile

SNS Responds to the COVID-19 Pandemic

The Strategic National Stockpile's role is to supplement state and local supplies during public health emergencies. Many states have products stockpiled, as well. The supplies, medicines, and devices for life-saving care contained in the stockpile can be used as a short-term stopgap buffer when the immediate supply of adequate amounts of these materials may not be immediately available.

About the Stockpile | Sustaining the Stockpile | Stockpile Products



# We get URLs cited in Wikipedia via the EventStreamAPI



## Event Platform/EventStreams

< [Event Platform](#)  
(Redirected from [EventStreams](#))

**EventStreams** is a web service that exposes continuous streams of structured event data. It does so over HTTP using [chunked transfer encoding](#) following the [Server-Sent Events](#) (SSE) protocol. EventStreams can be consumed directly via HTTP, but is more commonly used via a client library.

EventStreams provides access to arbitrary streams of data, including [MediaWiki RecentChanges](#). It replaced [RCStream](#), and may in the future replace [irc.wikimedia.org](#). EventStreams is backed by [Apache Kafka](#).

*Note: Often 'SSE' and EventSource are used interchangeably. This document refers to SSE as the server-side protocol, and EventSource as the client-side interface.*

### Contents [hide]

- 1 [When not to use EventStreams](#)
- 2 [Examples](#)
  - 2.1 [JavaScript](#)
  - 2.2 [Python](#)
  - 2.3 [Command-line](#)
- 3 [API](#)
  - 3.1 [Stream selection](#)
  - 3.2 [Historical Consumption](#)
  - 3.3 [Response Format](#)
  - 3.4 [Filtering](#)
- 4 [Architecture](#)
  - 4.1 [SSE vs. WebSockets/Socket.IO](#)
  - 4.2 [KafkaSSE](#)
  - 4.3 [Kafka](#)
- 5 [WMF Administration](#)

```
HTML
<div class="container">
  <div class="events">
    <button type="button" value="Live">Live
  </div>
  <ul>
    <li>page create
    <li>page delete
    <li>page info change
    <li>page move
    <li>page properties change
    <li>page update
    <li>RecentChange
    <li>revision create
    <li>revision delete
    <li>revision restore
    <li>text
  </ul>
</div>

JS
const eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
eventSource.addEventListener("message", function(event) {
  // console.log(event.data);
});
eventSource.addEventListener("error", function(event) {
  // console.log(event);
});
const eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
const eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
const eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
const eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
```

Example client at [codepen.io/ottomata/pen/LYpPxj](https://codepen.io/ottomata/pen/LYpPxj).

```
HTML
<div class="container">
  <h1>RecentChange stats</h1>
  <div class="stats">
    <table>
      <tr>
        <td>All wikis</td>
        <td>www.wikidata.org</td>
        <td>en.wikipedia.org</td>
      </tr>
      <tr>
        <td>Total: 6,007</td>
        <td>Total: 2,921</td>
        <td>Total: 971</td>
      </tr>
      <tr>
        <td>First record: 23 Aug '20</td>
        <td>First record: 18 Aug '20</td>
        <td>First record: 2 Aug '21</td>
      </tr>
      <tr>
        <td>Pages involved: 2,269 (page 1,277)</td>
        <td>Pages involved: 584 (page 372)</td>
        <td>Pages involved: 183 (page 100)</td>
      </tr>
      <tr>
        <td>Pages moved: 404 (page 200)</td>
        <td>Pages moved: 44 (page 23)</td>
        <td>Pages moved: 14 (page 8)</td>
      </tr>
    </table>
  </div>
  <div class="social">
    <a href="#">
    <a href="#">
    <a href="#">
  </div>
</div>

JS
function init() {
  var eventSource = new EventSource(EventStreamsURL + "/events/" + stream);
  var errorNode = document.getElementById("error-messages");
  var successNode = document.getElementById("success-messages");
  eventSource.addEventListener("message", function(event) {
    // console.log(event);
  });
  eventSource.addEventListener("error", function(event) {
    // console.log(event);
  });
  eventSource.addEventListener("success", function(event) {
    // console.log(event);
  });
}
```


## RecentChange stats from EventStreams

All wikis	www.wikidata.org	en.wikipedia.org
Total: 6,007	Total: 2,921	Total: 971
First record: 23 Aug '20	First record: 18 Aug '20	First record: 2 Aug '21
Pages involved: 2,269 (page 1,277)	Pages involved: 584 (page 372)	Pages involved: 183 (page 100)
Pages moved: 404 (page 200)	Pages moved: 44 (page 23)	Pages moved: 14 (page 8)



# We Archive Millions of Wikipedia Sourced URLs/day



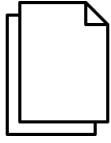
**cawg**   top space dash dump scan reduce viz

From 2020   To 2023   Set

Days 1224 collection  ☆ Meta-counts 1146/1220 (94%) ...

wikipedia-eventstream						
	date	items	count	size	B/uri	
1	<a href="#">2023-08-07</a> ☆	7	1,298,950 0%	82,731,858,081 0%	63.7KB	
2	<a href="#">2023-08-06</a> ☆	8	1,410,651 0%	81,116,676,520 0%	57.5KB	
3	<a href="#">2023-08-05</a> ☆	6	824,281 0%	60,862,226,293 0%	73.8KB	
4	<a href="#">2023-08-04</a> ☆	6	960,376 0%	62,092,605,426 0%	64.7KB	
5	<a href="#">2023-08-03</a> ☆	7	1,076,207 0%	82,631,879,715 0%	76.8KB	
6	<a href="#">2023-08-02</a> ☆	6	890,475 0%	81,443,293,754 0%	91.5KB	
7	<a href="#">2023-08-01</a> ☆	8	1,070,500 0%	81,339,625,595 0%	76.0KB	
8	<a href="#">2023-07-31</a> ☆	8	1,405,654 0%	83,193,245,880 0%	59.2KB	
9	<a href="#">2023-07-30</a> ☆	12	2,305,861 0%	132,761,583,245 0%	57.6KB	
10	<a href="#">2023-07-29</a> ☆	9	1,390,846 0%	93,290,117,260 0%	67.1KB	

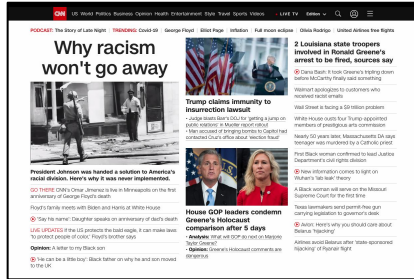
ALL collections					
	☆	items	count	size	B/uri
1	☆	5,934	8.6B	57.7TB	6.7KB
2	☆	6,337	9.4B	61.6TB	6.5KB
3	☆	6,241	9.6B	61.4TB	6.4KB
4	☆	6,863	7.5B	65.5TB	8.8KB
5	☆	5,741	3.4B	55.0TB	16.2KB
6	☆	5,888	8.2B	56.1TB	6.8KB
7	☆	4,816	6.0B	45.7TB	7.6KB
8	☆	4,317	3.1B	40.4TB	13.0KB
9	☆	4,253	2.8B	37.6TB	13.2KB
10	☆	4,372	2.4B	41.3TB	17.1KB



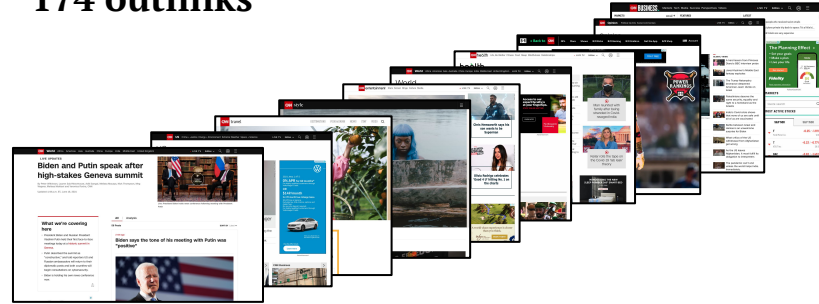
# How 1 seed URL can result in 31,259 Archived URLs



1 seed page



174 outlinks



338 embeds



30,746 outlink embeds




Results from cnn.com on June 17, 2021





# We also Archive Millions of Outlinked URLs



**cawg**   top space dash dump scan reduce viz

From 2020   To 2023

Days 1224 collection  ☆ Meta-counts 100%

wikipedia-eventstream-outlinks							
	date	items	count		size		B/uri
1	<a href="#">2023-08-07</a> ★	46	6,846,451	0%	467,531,356,333	1%	68.3KB
2	<a href="#">2023-08-06</a> ★	62	9,797,050	0%	639,846,297,269	1%	65.3KB
3	<a href="#">2023-08-05</a> ★	53	7,779,631	0%	561,287,474,335	1%	72.1KB
4	<a href="#">2023-08-04</a> ★	41	5,246,811	0%	416,772,334,461	1%	79.4KB
5	<a href="#">2023-08-03</a> ★	48	5,908,604	0%	496,407,096,697	1%	84.0KB
6	<a href="#">2023-08-02</a> ★	52	6,007,106	0%	544,033,016,699	1%	90.6KB
7	<a href="#">2023-08-01</a> ★	64	6,799,250	0%	661,515,880,081	1%	97.3KB
8	<a href="#">2023-07-31</a> ★	64	8,472,577	0%	672,326,227,733	2%	79.4KB
9	<a href="#">2023-07-30</a> ★	95	15,519,149	1%	973,588,838,302	3%	62.7KB
10	<a href="#">2023-07-29</a> ★	64	9,759,397	0%	673,523,724,314	2%	69.0KB

ALL collections						
	★	items	count	size	B/uri	
1	★	5,934	8.6B	57.7TB	6.7KB	
2	★	6,337	9.4B	61.6TB	6.5KB	
3	★	6,241	9.6B	61.4TB	6.4KB	
4	★	6,863	7.5B	65.5TB	8.8KB	
5	★	5,741	3.4B	55.0TB	16.2KB	
6	★	5,888	8.2B	56.1TB	6.8KB	
7	★	4,816	6.0B	45.7TB	7.6KB	
8	★	4,317	3.1B	40.4TB	13.0KB	
9	★	4,253	2.8B	37.6TB	13.2KB	
10	★	4,372	2.4B	41.3TB	17.1KB	



# An Interactive Page from the 117th Congress



Internet Archive  
Wayback Machine  
565 captures  
31 Dec 2022 - 17 Aug

Interactive Attack Timeline | Jan 6th Report

https://january6report.house.gov/timeline/

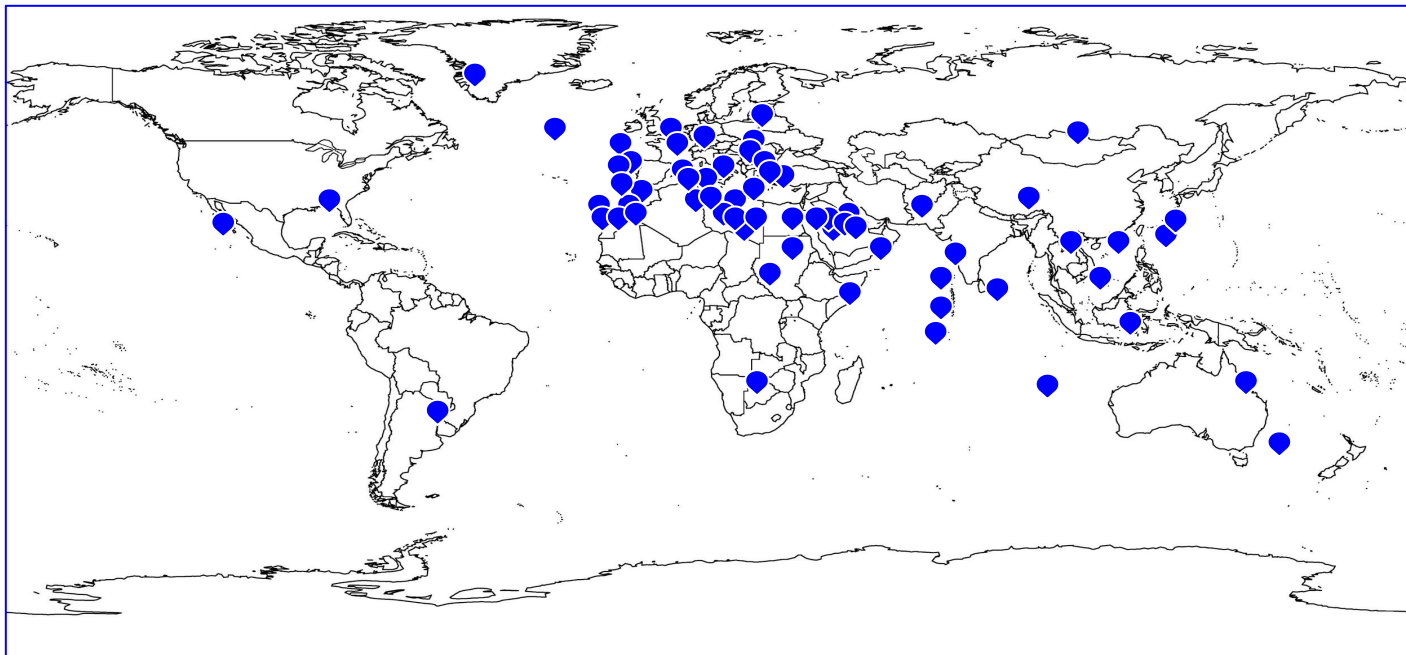
Go DEC JAN 01 2023 FEB

2. West Plaza  
3. Senate Wing Door and Window  
4. East Plaza of Capitol  
5. Rotunda  
6. House Wing  
7. Senate Wing

10:00 am 11:30 pm



IABot has run on **215** out of **334**  
Wikipedia language editions, and **58** additional wikis





# Progress Fixing Broken Links

- Fixed more than **19 million** formerly broken links across 305 Wikipedia languages
- More than **8 million** fixed in the last 2 years alone!



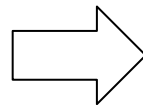




# Better World Books



The Internet Archive creates a wish list of books  
based on Wikipedia citations



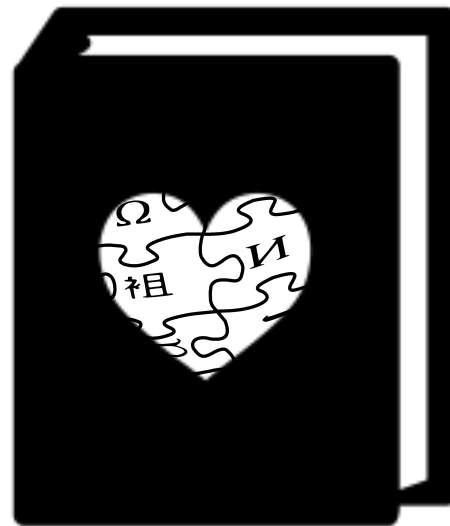


# Progress Adding Links to Books



More than **1.8 million links** to books from [archive.org](https://archive.org) across have been added to **~60 Wikipedia language editions**

(1 million added by our software and 800,000 added by Wikipedia Editors)





# Progress Adding Links to Papers



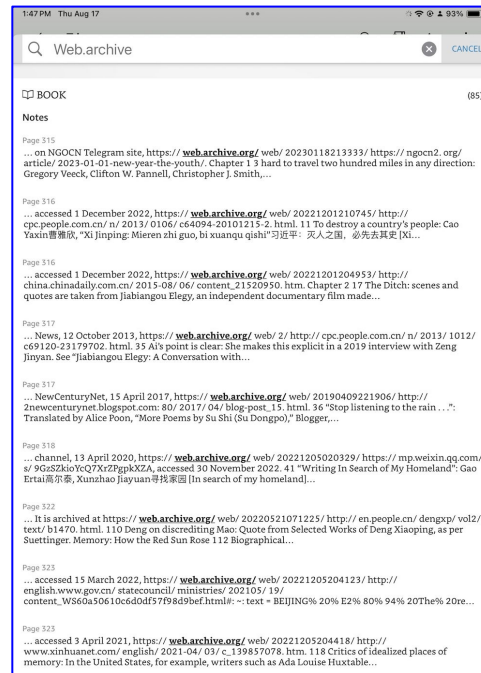
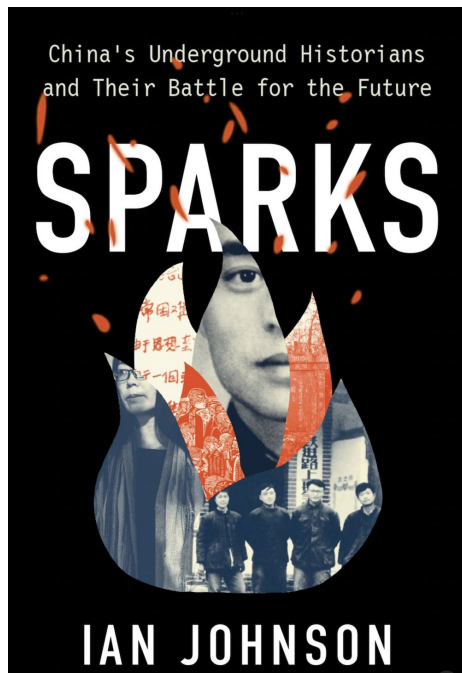
Our software has added more than **141,000** links to academic papers and articles across **~60 Wikipedia language** editions







# 85 Wayback Machine Links in this book







# Next Up: Finding/Fixing “Soft 404s”

The screenshot shows the Yoast SEO website interface. At the top, the logo reads "yoast SEO for everyone". To the right are links for "MyYoast", "Cart", "USD (\$)", and a search bar. Below is a navigation menu with buttons for "Home", "Yoast SEO plugin", "Other plugins", "SEO training", "SEO blog", and "Support". The main content area displays a 404 error message: "Home » Error 404: Page not found". The message says "Sorry, this page could not be found." and provides instructions: "Sorry! The page you are looking for doesn't exist or no longer exists. Some options for how to find it: 1. If you typed in a URL... make sure the spelling, cApitALiZaTiOn, and punctuation are correct. Then try reloading the page. 2. Try and find it in the [sitemap](#). 3. Start over again at the [homepage](#). 4. Search for it:" followed by a search input field with the placeholder text "Search" and a search icon.

yoast SEO for everyone

MyYoast Cart USD (\$) Search...

Home Yoast SEO plugin Other plugins SEO training SEO blog Support

Home » Error 404: Page not found

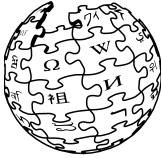
## Sorry, this page could not be found.

Sorry! The page you are looking for doesn't exist or no longer exists. Some options for how to find it:

1. If you typed in a URL... make sure the spelling, cApitALiZaTiOn, and punctuation are correct. Then try reloading the page.
2. Try and find it in the [sitemap](#).
3. Start over again at the [homepage](#).
4. Search for it:

Search


Q



# Publishers Lawsuit (blog.archive.org)



## UNIVERSAL ACCESS TO ALL KNOWLEDGE



Blog Announcements archive.org About Events Developers Donate

### What the Hachette v. Internet Archive Decision Means for Our Library

Posted on August 17, 2023 by [Chris Freeland](#)

*Our library is still strong, growing, and serving millions of patrons. But the publishers' attack on basic library practices continues.*

**Martin Luther King Jr.**  
From Wikipedia, the free encyclopedia

"Martin Luther King" and "MLK" redirect here.

Other names:

- 121. <sup>a</sup> Cohen & Taylor 2000, p. 347.
- 122. <sup>a</sup> Cohen & Taylor 2000, p. 416.
- 123. <sup>a</sup> Fairclough, Adam (1987). *To Reclaim the Soul of America: The Southern Christian Leadership Conference & Martin Luther King Jr.*. University of Georgia Press. ISBN 0-8203-2346-2. [\[Read on Archive.org\]](#)

**Preview**

Some pages are omitted from this book preview.

... about in Chicago

... most persons and perhaps even less of the "The book he should be read to "reverse racial attitudes" of common members, "with a heavy heart," he urged them on to do "the necessary" before "the dawn of the United States had changed for good. In March, he had vowed that "he'd stick it out the rest of the year. The six African U.S. leaders were set on an evening rally and their ordering Civil's plea - "The sudden knowledge of what happened was not within the MLK had failed to get the attention" the Board affirmed that "those who would expose their intellectual discrimination demonstrate their personal savings. When word came that a press would be coming, however, Paul and Ruby turned Civil's request. Archives" ... not to abandon us now. "Ruby urged a persistent Civil to dropping the manuscript. "What if I don't write up the United States in time," he asked. MLK, we've got news for the man. If the history doesn't get us back up to China, let him give the field." The history needs his statement. Read full the handbook and the long wrapping around the office of the Chicago for Two days later, Bevel, Jackson, and Ruby took over the full of the who, see, see, Bevel. When local members requested, they decided to hold simultaneous meetings at Park and Johnson Park on August 14. No one had but Coretta, but the noise, and multiple members, who joined to the house, served equally well. About one-third had part to be sure a tight press. Although there was a small throng of books and books, the most were

Linking references in Wikipedia to the cited page in our library.

#### Latest Posts

- [What the Hachette v. Internet Archive Decision Means for Our Library](#)
- [Internet Archive Responds to Recording Industry Lawsuit Targeting Obsolete Media](#)
- [Our Fight is Far From Over](#)
- [DWeb Fellows 2023: Lighting the Path Towards a Better Web](#)
- [Weaving Together the Story of Historic Lace Using the Internet Archive](#)

#### Upcoming Events

AUG		10:00 am - 11:00 am	
<b>24</b>	<b>Book Talk: Moving Theory Into Practice</b>		
OCT		10:00 am - 11:30 am	
<b>4</b>	<b>Library Leaders Forum: VIRTUAL</b>		
OCT		6:00 pm - 8:00 pm	