Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Dr. Michele C. Weigle
Professor, Department of Computer Science
Web Science and Digital Libraries (WS-DL) Group
Old Dominion University

Featuring work done in collaboration with

Web Science and Digital Libraries (WS-DL) Research Group at ODU

PhD Students and Alumni

Sawood Alam (PhD 2020)  
Mat Kelly (PhD 2019)  
Hany SalahEldeen (PhD 2015)

MS Alumni

Nauman Siddique (MS 2020)

REU Students (undergrad)

Lesley Frew  
Kritika Garg  
Himalsha Jayanetti  
Tarannum Zaki  
Haley Bragg  
Caleb Bradford

Dr. Michael L. Nelson

Funders

mweigle@odu.edu (@weiglemc) | ws-dl.cs.odu.edu (@WebSciDL) | April 19, 2023 | Lucy Family Institute for Data and Society
Computer scientists are toolsmiths

The Web holds our stories

ESPN.com  March 13, 2020

NCAA tournaments canceled over coronavirus
The NCAA has called off its men’s and women’s basketball tournaments as part of a complete cancellation of all remaining spring and winter championships.
Sh - ESPN News Services

What we know and don’t know after NCAA tournament cancellations

CNN.com  March 14, 2020

Trump declares national emergency

The declaration comes as there is only one US state left without coronavirus cases

Get Info: Send your stories | Podcast | Newsletter | Alerts

Trump says he takes no responsibility for early testing failures
Analysis: Trump’s appalling, blame-shifting briefing
See Oval’s response to Trump’s attempt to shake hands 😊
But webpages can change or disappear

ESPN.com  April 14, 2023

Debating biggest need vs. best value in Round 1: We made two picks at all 31 NFL draft slots
Fill a need? Best player available? We played out both strategies at all 31 first-round slots in the 2023 draft.

STACKING THE DRAFT'S TOP CORNERBACKS: HOW THEY SHUT DOWN WIDEOUTS

CNN.com  April 14, 2023

Pentagon document leak suspect charged

The 21-year-old Air National Guardsman appeared in federal court in Boston a day after his arrest at his Massachusetts home

Teixeira entered court in shackles; his hands were uncuffed
Man arrested in connection with intel leak was a loner fascinated
Maybe it's archived?

More than just saving a screenshot, web archives strive to capture and allow replay of the entire contents of a web page, including its source HTML and embedded images, stylesheets, and JavaScript source code.

Web archives are essential for studying recent history and culture

You can’t study the 1990s without web archives.

And historians aren’t ready...

Ian Milligan (University of Waterloo), https://twitter.com/NetPreserve/status/1141321443373920256
There are several public web archives

How do webpages get archived?

Anatomy of a Web Page

https://www.cs.odu.edu/~mweigle/
How do webpages get archived?

Web Crawling Process

Seed list

Frontier

Crawler

request and download resource

process resource

next URL

extracted URLs

resource headers and content

Storage (WARC files)

The Web
On-Demand Web Archiving

**Add Seeds**

Enter one seed URL per line below to add them to this collection.


**Access:** Private

**Frequency:** One-Time

**Seed Type:** One Page

- Standard
- One Page Plus External Links (One Page+)
- Standard Plus External Links (Standard+)

**Save Page Now**

https://web.archive.org

**New Capture**

URL to capture

- https://

Add to collection

- Public

Select browser

- Use Current Browser

**Session settings**

<table>
<thead>
<tr>
<th>browser</th>
<th>version</th>
<th>release</th>
<th>OS</th>
<th>capabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chrome</td>
<td>v76</td>
<td>2019-08-05</td>
<td>linux</td>
<td>flash, autoplot</td>
</tr>
<tr>
<td>Firefox</td>
<td>v68</td>
<td>2019-07-09</td>
<td>linux</td>
<td>flash, autoplot</td>
</tr>
</tbody>
</table>

**archive.today**

https://archive.today

**Conifer**

by RHIZOME.org

https://conifer.rhizome.org
Webrecorder and Conifer

https://guide.conifer.rhizome.org/


https://twitter.com/IlyaKreymer/status/1288193177279541248

Thanks @machawk1 for the generous mention! 😊

A well-deserved award! Several years before Webrecorder, WARCcreate really paved the way in providing an alternative, user-driven approach to small-scale web archiving!

https://twitter.com/IlyaKreymer/status/1288193177279541248
Query web archives for https://www.nd.edu/

https://web.archive.org/

https://archive.today/
List of Mementos for https://www.nd.edu/

http://web.archive.org/web/202300000000000*https://www.nd.edu

https://archive.is/https://www.nd.edu/
Navigating the Wayback Machine

https://web.archive.org/web/19990208021711/https://www.nd.edu/

Notre Dame Email Home Page

You can search:
- **X.500 Directory Service** (Find email addresses of Notre Dame Students, Faculty, and Staff)

You can administer your account:
- **Remove Forwarding Of Your ND Email Account** (All mail is sent to your ND mail account. You will need to configure a mail client like Netscape to read your mail.)
- **Forward Your ND Email** (All mail is sent to the address you specify, e.g., bob@aol.com. Do NOT use this to forward mail to yourself at nd.edu. Use the link directly above to receive mail at nd.edu)
- **Configure Vacation Auto-reply**
- **View IMAP quota information**

You can work with LISTSERV lists:
- **ND LISTSERV Home Page**

You can learn Eudora:
- **Eudora Tutorial** (Windows PC)
Navigating the Wayback Machine

https://web.archive.org/web/20010118210700/https://www.nd.edu/
Can also issue a prefix query to the Wayback Machine

http://web.archive.org/web/*/https://www.nd.edu/*

Lists all URLs archived that start with www.nd.edu or nd.edu
Querying Multiple Web Archives Using Memento Time Travel


https://timetravel.mementoweb.org/

http://timetravel.mementoweb.org/list/20140425043528/http://www.nd.edu/
Closest Mementos to April 25, 2014

Internet Archive - 20140425003923

archive.is - 20140421023111

Stanford - 20140512070512

Library of Congress - 20140307142751

National Library of Israel - 20140806155143

Arquivo.pt - 20140906175952
Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Using Web Archives to Document Social Movements - Practice
Egyptian Revolution (2011)

- Called the "Facebook Revolution" because of the large role that social media played in helping protesters to organize
- Study revealed that after 1 year, 10% of the resources *linked to* in tweets were no longer available
  - only 20% of the resources had been archived
- Percentage lost and percentage archived increased over time
- This held for multiple events between June 2009 - March 2012

Hany SalahEldeen. Losing My Revolution: A year after the Egyptian Revolution, 10% of the social media documentation is gone., https://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html
Documenting Ferguson (2014)

http://digital.wustl.edu/ferguson/

The movement never would've moved if we didn't have cell phones and Twitter. 2014. #Ferguson

https://twitter.com/deray/status/549442824752889856
Collecting Large Tweet Datasets

- Created via the Twitter API, using tools such as twarc
- Download JSON content of tweets, including tweet text and metadata

https://github.com/DocNow/twarc

Twitter's TOS only allows sharing large datasets as Tweet IDs

https://twitter.com/lucy_institute/status/1648672988525387777

tweet id

Turning tweet ID into tweet content (JSON format) requires "hydration" using Twitter's API
Ferguson Twitter Archive

https://catalog.docnow.io/datasets/20161224-ferguson-tweets/


28M tweet IDs
"ferguson"
2014-08-10 - 2014-08-27
2014-11-11 - 2014-12-08
Archiving #BlackLivesMatter (2016)

https://catalog.docnow.io/datasets/20171017-blacklivesmatter-2016/

<table>
<thead>
<tr>
<th>Title: BlackLivesMatter 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repository: Internet Archive</td>
</tr>
<tr>
<td>Creator(s): Ed Summers</td>
</tr>
<tr>
<td>Subjects: Blacklivesmatter, Activism</td>
</tr>
<tr>
<td>Dates: 01/29/2016 - 03/18/2017</td>
</tr>
<tr>
<td>Number of Tweets: 17,292,130</td>
</tr>
</tbody>
</table>

Description:
This dataset contains 17,292,130 tweet ids for tweets collected from the Twitter filter stream API for #blm and #blacklivesmatter between 2016-01-29 and 2017-03-18. The data was collected using the twarc utility. The files are broken into segments because of network connectivity problems that were encountered during data collection. So there are varying time gaps present between the files. Also when the hashtags were trending globally rate limits may have prevented some tweets from being streamed over the API.

17M tweet IDs
#blm, #blacklivesmatter
2016-01-29 - 2017-03-18
Web Archiving Tweets

- Web archives access tweets via the web (not Twitter's API)
- Can archive a single tweet (using the post URL) or a set of tweets from an account (using the account URL)
- Does not depend upon the Twitter API for replay
- Can view deleted tweets and tweets from deleted/suspended accounts
- Processing tweet content requires web scraping
- Archiving and later analysis occurs on a much smaller scale
#blacklivesmatter Web Archive

https://archive-it.org/collections/4783

1060 URLs, including 95 Twitter accounts, hashtags, and search results

started Aug 2014
COVID-19 (2020)

https://covid19.archive-it.org/

154 collections at Archive-It
South Carolina Dept of Education, Dedication to Education Page

January 14, 2021

today redirects to SC DOE's COVID Newsroom,

https://wayback.archive-it.org/4887/*/http://dedicationtoeducation.com/
#IStayHomeFor on Twitter (April 2020)

https://wayback.archive-it.org/4887/20200426034124/https://twitter.com/hashtag/IStayHomeFor/

April 26, 2020
#SongsOfComfort on Twitter (March 2020)


March 19, 2020
Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Using Web Archives to Document Disinformation - Practice
Journalists have been using web archives for a while now

"God bless you, Wayback Machine"
- Rachel Maddow, Dec 16, 2016

“... conference addressing, understanding and preventing homosexuality.”

Last Week Tonight, Mar 18, 2018
Internet Archive even tracks its mentions in news stories

Example of news story with links to the Wayback Machine

Taylor Lorenz, "Meet the woman behind Libs of TikTok", Washington Post, April 2022,
https://www.washingtonpost.com/technology/2022/04/19/libs-of-tiktok-right-wing-media/
Researchers submit webpages to Internet Archive for later reference

Election Integrity Partnership archived posts about the 2020 Election at Internet Archive so they could go back and write about them later.

Unfortunately, web archives can be used to spread removed content

Joan Donovan, "Covid hoaxes are using a loophole to stay alive—even after content is deleted", April 2020, https://www.technologyreview.com/2020/04/30/1000881/covid-hoaxes-zombie-content-wayback-machine-disinformation/
However, Internet Archive is adding labels to some of this content

Mark Graham, "Fact Checks and Context for Wayback Machine Pages", Oct 2020,
We need web archives to study original materials later, even if they contain misinformation.

“It’s not about trying to archive the stuff that’s true, but archive the conversation. All of that is what people are experiencing.”

-Mark Graham, Director of the Wayback Machine

https://www.ft.com/content/5be1f2ee-d60b-11e9-a0bd-ab8ec6435630

Using Web Archives to Find Deleted Tweets

2018-04-23: "Grampa, what's a deleted tweet?"
By Nauman Siddique - April 23, 2018


2019-04-17: Russell Westbrook, Shane Keisel, Fake Twitter Accounts, and Web Archives
By Michael L. Nelson - April 17, 2019


On March 11, 2019 in the NBA, the Utah Jazz hosted their Northwest Division rivals, the Oklahoma City Thunder. During the game, a Utah fan (Shane Keisel) and an Oklahoma City player (Russell Westbrook) engaged in a verbal exchange, with the player stating the fan was directing racist comments to him and the fan admitting to heckling but denying that his comments were racist. The event was well documented (see, for example, this Bleacher Report article), and the following day the fan received a lifetime ban from all events at the Vivint Smart Home Arena and the player received a $25k fine from the NBA.

Disclaimer: I have no knowledge of what the fan said during the game, nor do I have an opinion regarding the appropriateness of the respective penalties. My interest is that after the game, the fan gave at least one interview with a TV station reporter in which he exposed his identity. That set off a rapidly evolving series of events with both real and fake Twitter accounts, which we unraveled with the aid of multiple web archives. The initial analysis was performed by Justin Whitlock as a project in my CS 895 "Web Archiving Forensics" class; prior to Justin proposing it as a project topic, my only knowledge of this event was via the Daily Show.
"Did They Really Tweet That?"


Tarannum Zaki, Disinformation Spread on Social Media through Screenshot Sharing: Dataset Description, https://ws-dl.blogspot.com/2022/12/2022-12-12-disinformation-spread-on.html
Find All Archived Tweets for a User at Internet Archive


http://web.archive.org/web/20211123000955/https://twitter.com/weiglemc/status/1003997508572471296
Why do journalists use web archives?

View unavailable page
- Deletion - page
- Site unavailable on live web

View change over time
- Content evolution
- Deletion - content

Investigate misinformation/hacking
- Dispute fabricated screenshot
- View historical version

Other
- Website lifespan

Searching Web Archives for Changed Text

- Only a few web archives offer full-text search.
- None (yet) allow you to search for what's changed.

<table>
<thead>
<tr>
<th>Search results for added phrase: non cloth</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumFound: 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>Novel Coronavirus (COVID-19)</th>
<th>Department of Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>url</td>
<td><a href="https://coronavirus.health.ny.gov/node/1">https://coronavirus.health.ny.gov/node/1</a></td>
<td></td>
</tr>
<tr>
<td>addition</td>
<td>2022-03-09 13:12:20</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>Novel Coronavirus (COVID-19)</th>
<th>Department of Health</th>
</tr>
</thead>
<tbody>
<tr>
<td>url</td>
<td><a href="https://coronavirus.health.ny.gov/home">https://coronavirus.health.ny.gov/home</a></td>
<td></td>
</tr>
<tr>
<td>addition</td>
<td>2022-02-11 21:20:50</td>
<td></td>
</tr>
</tbody>
</table>

- All travelers, and , should continue to follow all CDC travel requirements requirements.
- New Yorkers who don't feel well should stay home.
- Wear masks in public indoor space and wash hands regularly.

Last memento before the addition: https://wayback.archive-it.org/4887/20220203045021/https://coronavirus.health.ny.gov/home
First memento after the addition: https://wayback.archive-it.org/4887/20220211212050/https://coronavirus.health.ny.gov/home


deletion of the statement about George Santos's work history

George Anthony was then offered an exciting opportunity with Goldman Sachs but what he thought would be the pinnacle of his career was not as fulfilling as he had anticipated. George joined LinkBridge Investors, where he quickly advanced to become one of the youngest vice presidents in the industry.

George's work ethic comes from his mother, who came from nothing, but worked her way up to be the first female executive at a major financial institution. On September 11, 2001, George's mother was in her office in the South Tower. She survived the horrific events of that day, but unfortunately passed away a few years later.

Following in his mother's footsteps, George has a long history of philanthropic work across New York, and around the country. He has worked tirelessly with animal rescue efforts in the Tri-State area; and supports other organizations that help at-risk children and America's veterans.

George graduated from Baruch College with a bachelor's degree in economics and finance. He is fluent in Portuguese and Spanish. George lives on Long Island with his Husband, Matt, and their 4 dogs.

deletion of the statement about George Santos's grandparents being Jewish

George's grandparents fled Jewish persecution in Ukraine, settled in Belgium, and again fled persecution during WWII. They were able to settle in Brazil, where his mother was born. His father, who comes from Angolan roots, was also born in Brazil. Both his mother and father legally immigrated to the United States in search of the American dream, where they began their family on the foundations of life, liberty, and the pursuit of happiness.

George's drive, commitment and determination were inspired by the legacy set and left by his mother. She gave up everything to provide for George and her family. She delivered the American dream to George, a debt that he wants to repay to the rest of his neighbors and constituents. George's mother was in her office in the South Tower on September 11, 2001, when the horrific events of that day unfolded. She survived the tragic events on September 11th, but she passed away a few years later when she lost her battle to cancer.

The American dream that George grew up living is now under an all-out attack from the Radical Left. Inflation is out of control and crime is rampant in the streets. New York City and Long Island used to be the gold standard for quality of life in America – now New Yorkers have watched their communities
Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Using Web Archives to Document Social Movements and Disinformation - Ethics
Documenting the Now

- Emerged in 2016, in response to efforts to document social movements that were being recorded on social media
- Provides tools and best practices for archivists and researchers

https://www.docnow.io/

Documenting the Now develops open source tools and community-centered practices that support the ethical collection, use, and preservation of publicly available content shared on web and social media.

Documenting the Now responds to the public’s use of social media for chronicling historically significant events as well as demand from scholars, students, and archivists, among others, seeking a user-friendly means of collecting and preserving this type of digital content. Documenting the Now has a strong commitment to prioritizing ethical practices when working with social media content, especially in terms of collection and long-term preservation. This commitment extends to Twitter’s notion of honoring user intent and the rights of content creators. Documenting the Now is a core program of Shift Collective. We are extremely grateful for funding from the Mellon Foundation and technical support from the Princeton University Library.
Goal was to bring together activists, librarians, journalists, archivists, scholars, developers, and designers

https://eaw.rhizome.org/

Who owns a user's posts on a social media platform?

- Users lack awareness of how social media platforms use their data and how that data can be used by third parties (including researchers and journalists)
- Survey found that a significant number of Twitter users were unaware their tweets could be used by third parties without consent (Fiesler and Proferes, 2018)


How to handle potential disinformation or hate speech?

- Malicious actors can pretend to be regular users and spread disinformation
  - @Blacktivist Twitter account, created by Russian IRA, tweeted prolifically in support of BLM, captured as part of #blacklivesmatter research

- Amplification of disinformation and hate speech is a concern
  - Duke Univ opted to not digitize a collection from the Southern Poverty Law Center to avoid recirculating hate materials


How to be inclusive without causing harm?

- Increased potential for harm for members of marginalized communities, especially when participating in activities such as protests or other acts of civil disobedience

- Activists in marginalized groups can face disproportionate harm from surveillance
  - researcher collecting tweets related to Ferguson was contacted for data by a company that provides "situational awareness" tools to law enforcement


Ethical Conundrums and Budgeting

**Ethical conundrums**

- **Ethical murkiness**: What social media to proactively collect or not collect?
  - #MeToo: Important social movement, but women are sharing personal and traumatic experiences.
  - #Parkland / #NeverAgain: Important event and social movement, but involves minors.
- **Each of these areas are important, but involve vulnerable individuals / communities or sensitive topics.**
- **Every decision not to collect results in a gap in historical record.**


[https://twitter.com/ianmilligan1/status/977267978772320257](https://twitter.com/ianmilligan1/status/977267978772320257)

[https://twitter.com/ianmilligan1/status/977262798488076290](https://twitter.com/ianmilligan1/status/977262798488076290)
Recommendations for Preserving Web/Social Media Content in Ethical Ways

● Engage and work with the communities being documented.

● Social media content should not be the only documentation collected.
  ○ "Simply telling the story of what happens in social media is not enough, but it can be a beginning."

● Follow social media platforms' TOS only where it lines up with the values of the communities being documented.
  ○ "What is legal is not always ethical, and what is ethical is not always legal."

● Apply traditional archival practices (appraisal, collection development, donor relations) to social media and web materials when possible.

Representation in Web Archives

- "For digital collections, who gets represented is closely tied with who:
  - writes the software
  - builds the tools
  - produces the technical standards
  - provides the funding"

- "The work we do as archivists, as librarians, as digital preservationists, have real consequences for marginalized people because who is remembered and how they’re remembered dictates who gets violence perpetrated against them."

- "Neutrality is a threat to the legacies of marginalized people and by extension their lives."

Archiving the Black Web

- Goal is to establish a "more equitable and accessible web archiving practice" to document the Black experience online
- Received $2.5M grant from the Mellon Foundation
- Develop training program in web archiving aimed at documenting Black life, history, and culture

https://archivingtheblackweb.org/
Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Using Web Archives to Document Social Movements and Disinformation - Challenges: Twitter
We spent the Fall of 2020 studying archived Twitter


based on a series of blog posts:


The live web keeps changing, so web archives must adjust to keep up

2009


2013


2017 (Old UI)


2017 (Old UI)


2020 (New UI)


mweigle@odu.edu (@weiglemc) | ws-dl.cs.odu.edu (@WebSciDL) | April 19, 2023 | Lucy Family Institute for Data and Society
@realDonaldTrump Twitter account is well archived

http://web.archive.org/web/20200701000000*/https://twitter.com/realDonaldTrump

Factba.se
https://factba.se/trump/

The Trump Archive
https://www.thetrumparchive.com/

http://web.archive.org/web/20200701000000*/https://twitter.com/realDonaldTrump

Garg et al., ACM/IEEE JCDL 2021
Used @realDonaldTrump to study the impact of Twitter UI change on web archives

- **2020-05-01**: Twitter stopped supporting its old UI
- **2021-01-08**: Permanent Suspension of @realDonaldTrump
- **2022-11-19**: Elon Musk brings Donald Trump back on Twitter

Collected ~8 months of archived data of @realDonaldTrump to quantify the impact of the change

No content on live web for ~2 years as account was suspended

https://blog.twitter.com/en_us/topics/company/2020/suspension
https://en.wikipedia.org/wiki/Acquisition_of_Twitter_by_Elon_Musk
Tweets in old UI are embedded in HTML / New UI requires JSON requests

Old UI

New UI

20 tweets and Twitter bio and are embedded in the base HTML
Content populated with follow-up XHR requests
base HTML contains only a skeleton, and all page sections are served dynamically through API JSON responses

(https://api.twitter.com/2/timeline/profile/25073877.json?..)
New UI mementos may replay pages that never existed on live web

Each section could have a different datetime
Twitter introduced labels in May 2020

Were being widely used by November 2020

https://ws-dl.blogspot.com/2020/12/2020-12-08-twitter-added-labels-on-its.html
We categorized labels as “Fact-check” and “Violated Twitter Rules” (VTR)

**Fact-check Labels**

Donald J. Trump
@realDonaldTrump

There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent. Mail boxes will be robbed, ballots will be forged & even illegally printed out & fraudulently signed. The Governor of California is sending Ballots to millions of people, anyone.....

Get the facts about mail-in ballots

8:17 AM · May 26, 2020 · Twitter for iPhone

36.1K Retweets 11.7K Quote Tweets 125.3K Likes

https://twitter.com/realDonaldTrump/status/1265255835124539392

**Violated Twitter Rules Labels**

Donald J. Trump
@realDonaldTrump

This Tweet violated the Twitter Rules about spreading misleading and potentially harmful information related to COVID-19. However, Twitter has determined that it may be in the public’s interest for the Tweet to remain accessible. Learn more

Flu season is coming up! Many people every year, sometimes over 100,000, and despite the Vaccine, die from the Flu. Are we going to close down our Country? No, we have learned to live with it, just like we are learning to live with Covid, in most populations far less lethal!!!

8:03 AM · Oct 6, 2020 · Twitter for iPhone

View Quote Tweets

https://twitter.com/realDonaldTrump/status/13134498441392961

Garg et al., ACM/IEEE JCDL 2021
Most web archives continued to archive the old UI, but labels were only included in Twitter’s new UI

[Image of old and new Twitter UI comparison]

Garg et al., ACM/IEEE JCDL 2021
Twitter added VTR label to old UI by at least August 26, 2020

https://ws-dl.blogspot.com/2020/12/2020-12-08-twitter-added-labels-on-its.html
But, the “Fact-check” label is still not a part of the old Twitter UI


Today, the “Fact-check” label is no longer on the live web either

Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

Using Web Archives to Document Social Movements and Disinformation - Challenges: Instagram
Instagram was a significant front in the IRA’s influence operation

- Russian Internet Research Agency (IRA)
- Manipulated social media narratives to influence American culture and politics in run-up to 2016 Presidential election
- Employed 1000+ to engage in influence operations, budget of over $25M
- Analysis commissioned by the US Senate Select Committee on Intelligence

However, Instagram is not as studied as Twitter or Facebook

<table>
<thead>
<tr>
<th>Twitter</th>
<th>Instagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>450 million monthly active users</td>
<td>2 billion monthly active users</td>
</tr>
<tr>
<td>8 million hits in Google Scholar</td>
<td>3 million hits in Google Scholar</td>
</tr>
<tr>
<td>Open (for now) API allows for searching and</td>
<td>API is focused more on allowing creation of</td>
</tr>
<tr>
<td>downloading public tweet content</td>
<td>content and monitoring engagement</td>
</tr>
<tr>
<td>Sharing based on retweets</td>
<td>No native sharing feature (though there are</td>
</tr>
<tr>
<td></td>
<td>methods for &quot;regrams&quot;)</td>
</tr>
<tr>
<td>Public posts and accounts viewable to not</td>
<td>Public posts and accounts limited for not</td>
</tr>
<tr>
<td>logged in users and bots</td>
<td>logged in users and bots</td>
</tr>
</tbody>
</table>
Instagram has banned accounts tied to misinformation

Facebook takes down Ukraine disinformation network and bans Russian-backed media

Meta says network ran websites posing as independent news entities and created fake personas

Facebook and Instagram have taken down a disinformation network targeting people in Ukraine, as their owner announced it was blocking access to the Russian state media outlets RT and Sputnik across the European Union.


Nearly 1,000 Facebook, Instagram accounts deactivated over COVID-19 misinformation

Meta deleted nearly 1,000 Instagram and Facebook accounts, groups in several countries last month, most of which were found to belong to a sprawling network from China trying to spread COVID-19 misinformation.

The deactivated accounts, originating in Palestine, Poland, Belarus and China, violated Meta's Coordinated Inauthentic Behavior (CIB) policy, which is when users manipulate public debate for a strategic goal where fake accounts are central to the operation.


Instagram meme account with 14 million followers banned for coronavirus scam

No, PayPal is not sending people $1,000.

By Matt Binder on March 30, 2020


RFK Jr. kicked off Instagram for vaccine misinformation

Instagram on Wednesday banned Robert F. Kennedy Jr., son of former presidential candidate Robert Kennedy, from repeatedly posting misinformation about vaccine safety and COVID-19.

Kennedy Jr. has amassed a huge following on social media, where he frequently posts debunked or unproven claims about vaccines. He also uses his social media pages to post about large pharmaceutical firms and environmental health concerns.

For instance, Instagram has banned 10 of the "Disinformation Dozen"

1. Joseph Mercola
2. Robert F. Kennedy, Jr.
3. Ty and Charlene Bollinger
4. Sherri Tenpenny
5. Rizza Islam
6. Rashid Buttar
7. Erin Elizabeth
8. Sayer Ji
9. Kelly Brogan
10. Christiane Northrup
11. Ben Tapper
12. Kevin Jenkins

Just 12 People Are Behind Most Vaccine Hoaxes On Social Media, Research Shows
Updated May 14, 2021 • 11:48 AM ET

https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes

Majority of Covid misinformation came from 12 people, report finds

CCDH finds ‘disinformation dozen’ have combined following of 59 million people across multiple social media platforms

Center for Countering Digital Hate, "The Disinformation Dozen: Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers", 2021,
https://counterhate.com/research/the-disinformation-dozen/
You can't view banned accounts on the live web

instagram.com/robertfkennedyjr/

instagram.com/drbuttar/

instagram.com/drchristianenorthrup/
Need web archives to examine content before accounts banned

instagram.com/robertf kennedyjr/
598 archived IG account pages for RFK, Jr. in IA


mweigle@odu.edu (@weiglemc) | ws-dl.cs.odu.edu (@WebSciDL) | April 19, 2023 | Lucy Family Institute for Data and Society
But most of these just redirect to the Instagram Login Page

Bragg et al., ACM/IEEE JCDL 2023
Only 6 successful replays for instagram.com/RobertFKennedyJr

<table>
<thead>
<tr>
<th>Mementos</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redirects to Login Page</td>
<td>592</td>
</tr>
<tr>
<td>Replayed Account Pages</td>
<td>6</td>
</tr>
<tr>
<td>Total Mementos</td>
<td>598</td>
</tr>
</tbody>
</table>

Bragg et al., ACM/IEEE JCDL 2023
Even if archived, not all post images may be replayable

No loaded post images: 3

Partial loaded post images: 1

Complete loaded post images: 2

Bragg et al., ACM/IEEE JCDL 2023
Instagram has made it even harder to archive pages

2021, 2022: Low percentages of replayable mementos and number post images available in those replayable mementos

Bragg et al., ACM/IEEE JCDL 2023
Using Web Archives to Document Social Movements and Disinformation - Practice, Ethics, and Challenges

- Web archiving is an important tool for documenting events in our society.
- Web archives have been used to help document social movements, important cultural events, and disinformation on the web.
- Developers and archivists need to collaborate with activists and members of affected communities to ensure there is representation, but not harmful exposure.
- Changes in social media platforms and web technologies present multiple challenges for web archives.