Archiving Digital Marketing

Examining Preservation of Dynamic Content on the Web Through the Lens of Online Advertisements

Christopher Rauch, Alex H. Poole, Travis Reid, Michele C. Weigle, Michael L. Nelson, Faryaneh Poursardar, and **Mat Kelly**

mkelly@drexel.edu • @machawk1 • @machawk1@digipres.club

A COLLABORATION OF

September 19, 2024

DREXEL UNIVERSITY College of Computing & Informatics & OLD DOMINION UNIVERSITY®

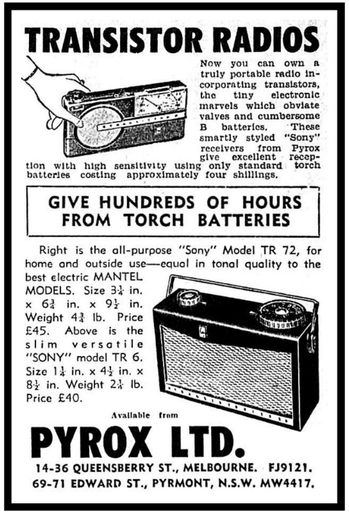
SUPPORTED BY

INSTITUTE of Museum and Library SERVICES

Slides: https://bit.ly/ipres2024

Why Preserve Online Advertisements

Advertisements are not just promotional tools but reflect and shape cultural and societal norms over time.



TRANSISTOR RADIOS

Now you can own a truly portable radio incorporating transistors, the tiny electronic marvels which obviate valves and cumbersome B batteries. These smartly styled "Sony" receivers from Pyrox give excellent reception with high sensitivity using only standard torch batteries costing approximately four shillings.

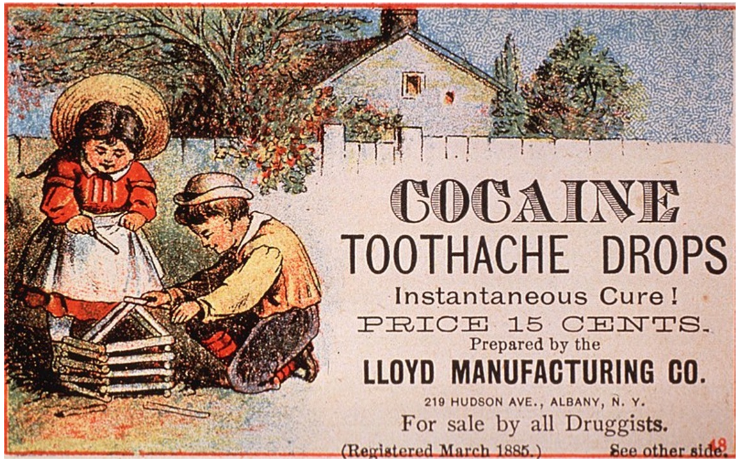
GIVE HUNDREDS OF HOURS FROM TORCH BATTERIES

Right is the all-purpose "Sony" Model TR 72, for home and outside use—equal in tonal quality to the best electric MANTEL MODELS. Size 3¼ in. x 6¾ in. x 9½ in. Weight 4¾ lb. Price £45. Above is the slim versatile "SONY" model TR 6. Size 1¼ in. x 4½ in. x 8½ in. Weight 2¼ lb. Price £40.

Available from

PYROX LTD.

14-36 QUEENSBERRY ST., MELBOURNE. FJ9121.
69-71 EDWARD ST., PYRMONT, N.S.W. MW4417.



COCAINE

TOOTHACHE DROPS

Instantaneous Cure!

PRICE 15 CENTS.

Prepared by the

LLOYD MANUFACTURING CO.

219 HUDSON AVE., ALBANY, N. Y.

For sale by all Druggists.

(Registered March 1885.) See other side.

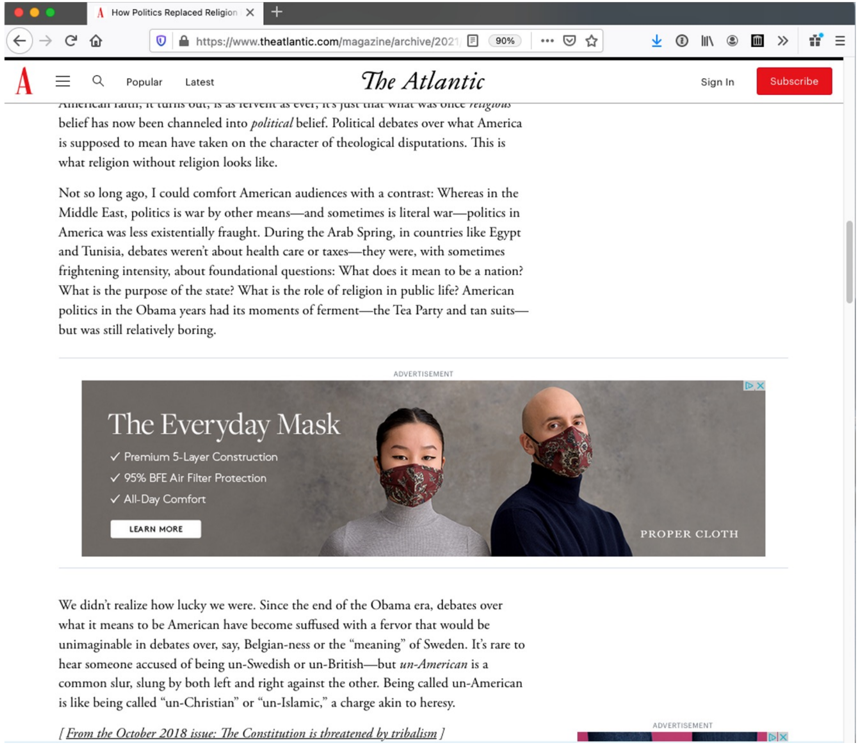
Source: Wikimedia Commons

Why Preserve Online Advertisements

ONLINE ADVERTISEMENTS

ARE DOCUMENTARY EVIDENCE

for understanding mid-1990s history onward


How Politics Replaced Religion
https://www.theatlantic.com/magazine/archive/2021
90%
A
Popular
Latest
The Atlantic
Sign In
Subscribe
ADVERTISEMENT
The Everyday Mask
✓ Premium 5-Layer Construction
✓ 95% BFE Air Filter Protection
✓ All-Day Comfort
LEARN MORE
PROPER CLOTH
ADVERTISEMENT

American faith, it turns out, is as fervent as ever; it's just that what was once *religious* belief has now been channeled into *political* belief. Political debates over what America is supposed to mean have taken on the character of theological disputations. This is what religion without religion looks like.

Not so long ago, I could comfort American audiences with a contrast: Whereas in the Middle East, politics is war by other means—and sometimes is literal war—politics in America was less existentially fraught. During the Arab Spring, in countries like Egypt and Tunisia, debates weren't about health care or taxes—they were, with sometimes frightening intensity, about foundational questions: What does it mean to be a nation? What is the purpose of the state? What is the role of religion in public life? American politics in the Obama years had its moments of ferment—the Tea Party and tan suits—but was still relatively boring.

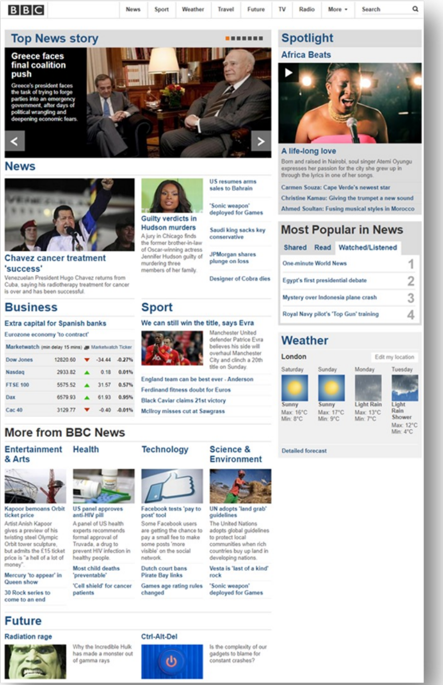
We didn't realize how lucky we were. Since the end of the Obama era, debates over what it means to be American have become suffused with a fervor that would be unimaginable in debates over, say, Belgian-ness or the "meaning" of Sweden. It's rare to hear someone accused of being un-Swedish or un-British—but *un-American* is a common slur, slung by both left and right against the other. Being called un-American is like being called "un-Christian" or "un-Islamic," a charge akin to heresy.

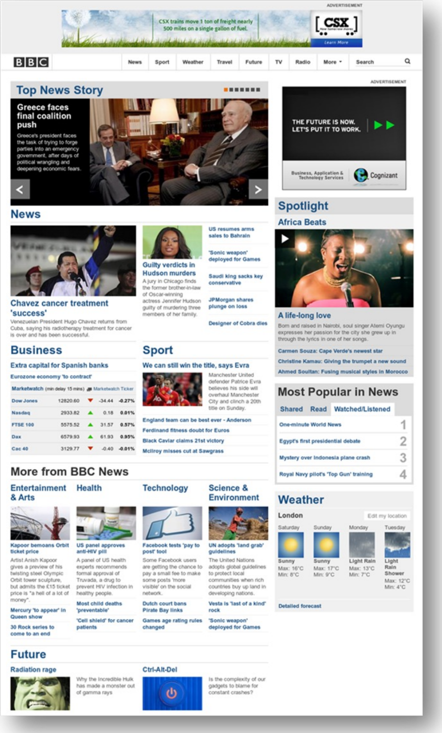
[From the October 2018 issue: The Constitution is threatened by tribalism]

Web Ads are a Rich Source for Historians

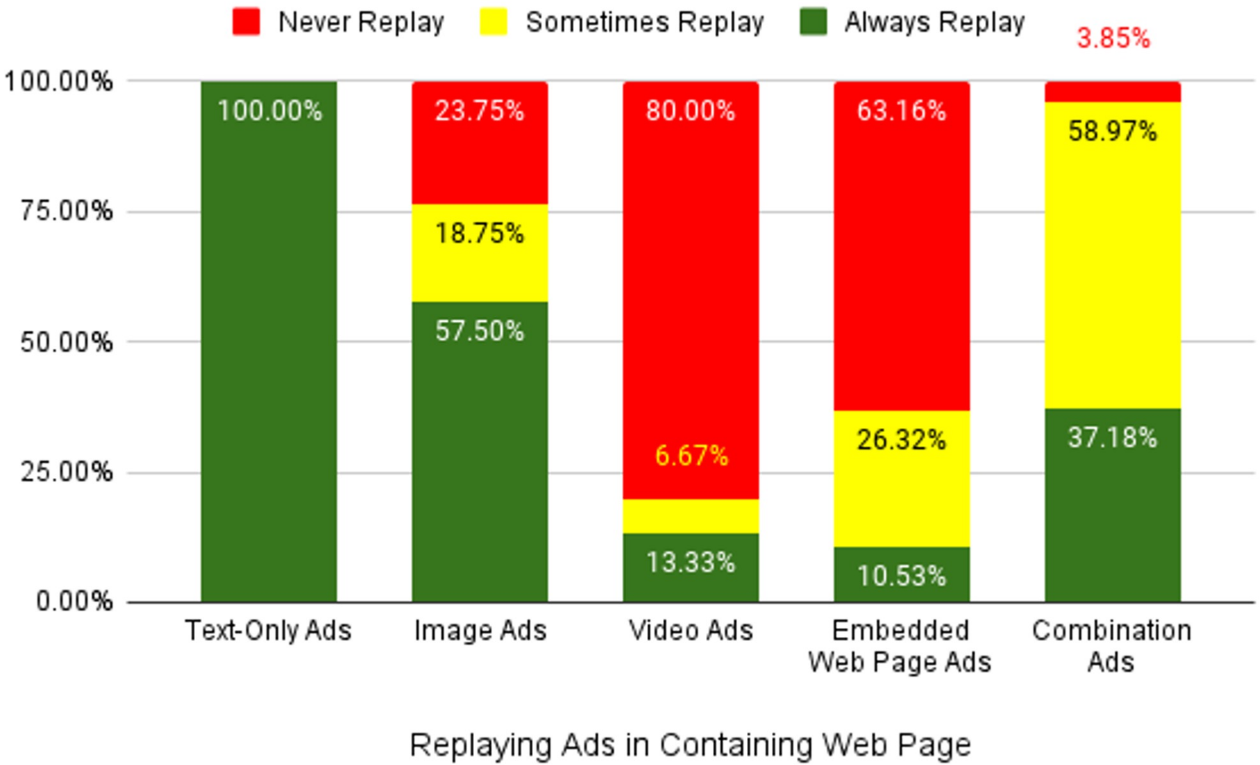
Archived ads enable scholars to:

- **EXPLORE** new questions
- **REVISIT** existing findings
- **ANALYZE** legal and social advertising practices.

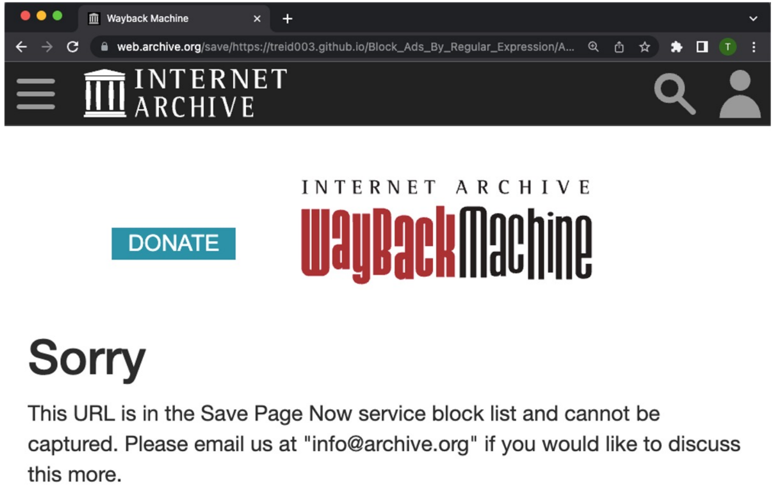

BBC
News Sport Weather Travel Future TV Radio More Search
Top News story
Greece faces final coalition push
Spotlight
Africa Beats
A life-long love
Carmen Souza: Cape Verde's newest star
Christine Kamau: Giving the trumpet a new sound
Ahmed Soultan: Fusing musical styles in Morocco
News
Chavez cancer treatment 'success'
Guilty verdicts in Hudson murders
US resumes arms sales to Bahrain
'Sonic weapon' deployed for Games
Saudi king sacks key conservative
JPMorgan shares plunge on loss
Designer of Cobra dies
Most Popular in News
Shared Read Watched/Listened
One-minute World News 1
Egypt's first presidential debate 2
Mystery over Indonesia plane crash 3
Royal Navy pilot's 'Top Gun' training 4
Business
Extra capital for Spanish banks
Eurozone economy 'to contract'
Sport
We can still win the title, says Evra
England team can be best ever - Anderson
Ferdinand fitness doubt for Euros
Black Caviar claims 21st victory
McIlroy misses cut at Sawgrass
Weather
London
More from BBC News
Entertainment & Arts
Health
Technology
Science & Environment
Kapoor bemoans Orbit ticket price
US panel approves anti-HIV pill
Facebook tests 'pay to post' tool
UN adopts 'land grab' guidelines
Future
Radiation rage
Ctrl-Alt-Del


ADVERTISEMENT
CSX trains move 1 ton of freight nearly 500 miles on a single gallon of fuel.
CSX
Learn More
BBC
News Sport Weather Travel Future TV Radio More Search
Top News Story
Greece faces final coalition push
ADVERTISEMENT
THE FUTURE IS NOW. LET'S PUT IT TO WORK.
Business, Application & Technology Services
Cognizant
News
Chavez cancer treatment 'success'
Guilty verdicts in Hudson murders
Spotlight
Africa Beats
A life-long love
Business
Sport
Most Popular in News
More from BBC News
Weather
Future

Challenges: Ephemerality

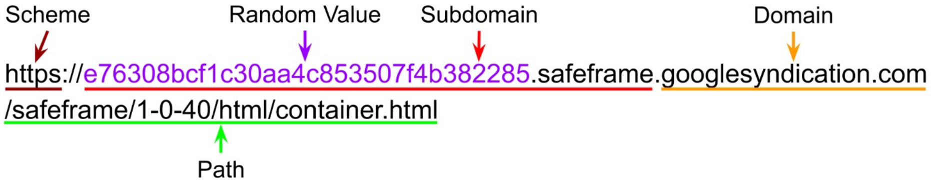


Challenges: Institutional Resistance

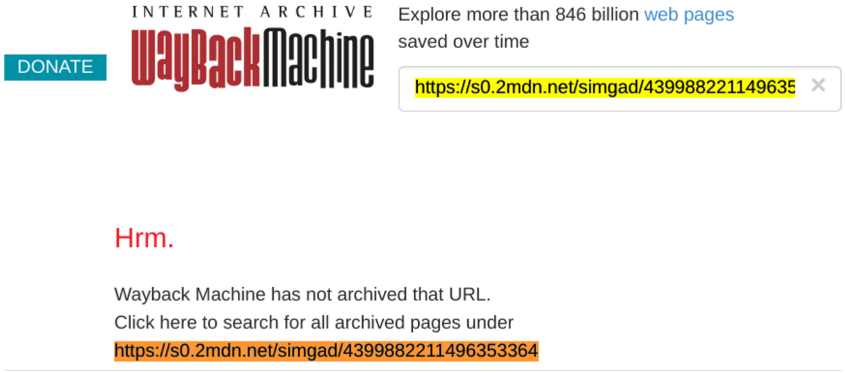

Wayback Machine
web.archive.org/save/https://treid003.github.io/Block_Ads_By_Regular_Expression/A...
INTERNET ARCHIVE
DONATE
INTERNET ARCHIVE
WaybackMachine

Sorry

This URL is in the Save Page Now service block list and cannot be captured. Please email us at "info@archive.org" if you would like to discuss this more.



https://e76308bcf1c30aa4c853507f4b382285.safeframe.googlesyndication.com/safeframe/1-0-40/html/container.html



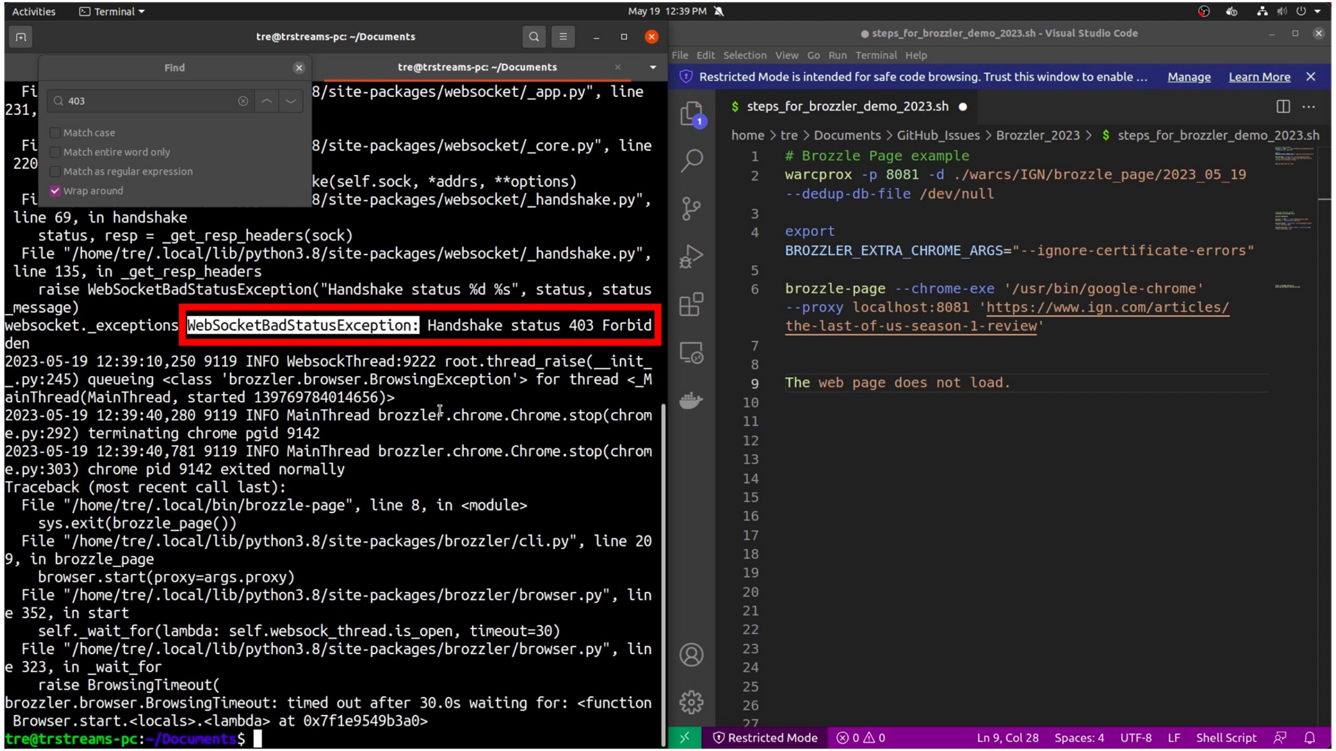
Explore more than 846 billion web pages saved over time

https://s0.2mdn.net/simgad/439988221149635 ✕

Hrm.

Wayback Machine has not archived that URL.
Click here to search for all archived pages under https://s0.2mdn.net/simgad/4399882211496353364

Enhanced User Interface Tools


Activities
Terminal
May 19 12:39 PM
tre@trstreams-pc: ~/Documents
Find
403
□ Match case
□ Match entire word only
□ Match as regular expression
☑ Wrap around
tre@trstreams-pc: ~/Documents
steps_for_brozzler_demo_2023.sh - Visual Studio Code
File Edit Selection View Go Run Terminal Help
Manage
Learn More
home > tre > Documents > GitHub_Issues > Brozzler_2023 > steps_for_brozzler_demo_2023.sh
Restricted Mode
⊗ 0 ⚠ 0
Ln 9, Col 28
Spaces: 4
UTF-8
LF
Shell Script

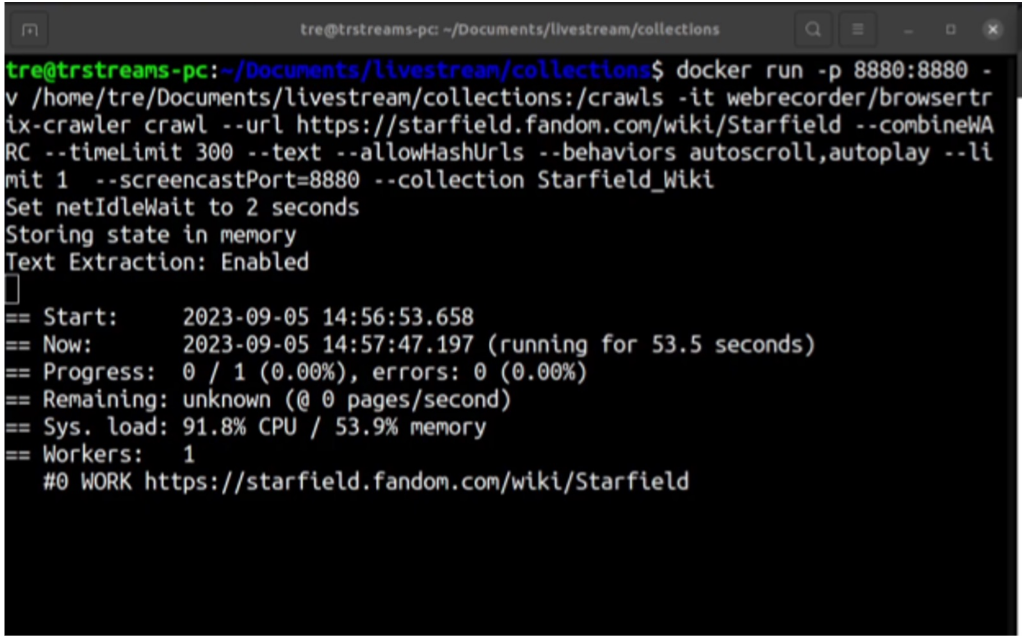
```
  Fi                                    8/site-packages/websocket/_app.py", line
231,

  Fi                                    8/site-packages/websocket/_core.py", line
 220

                                        ke(self.sock, *addrs, **options)
  Fi                                    8/site-packages/websocket/_handshake.py",
 line 69, in handshake
    status, resp = _get_resp_headers(sock)
  File "/home/tre/.local/lib/python3.8/site-packages/websocket/_handshake.py",
 line 135, in _get_resp_headers
    raise WebSocketBadStatusException("Handshake status %d %s", status, status
_message)
websocket._exceptions.WebSocketBadStatusException: Handshake status 403 Forbid
den
2023-05-19 12:39:10,250 9119 INFO WebsockThread:9222 root.thread_raise(__init_
_.py:245) queueing <class 'brozzler.browser.BrowsingException'> for thread <_M
ainThread(MainThread, started 139769784014656)>
2023-05-19 12:39:40,280 9119 INFO MainThread brozzler.chrome.Chrome.stop(chrom
e.py:292) terminating chrome pgid 9142
2023-05-19 12:39:40,781 9119 INFO MainThread brozzler.chrome.Chrome.stop(chrom
e.py:303) chrome pid 9142 exited normally
Traceback (most recent call last):
  File "/home/tre/.local/bin/brozzle-page", line 8, in <module>
    sys.exit(brozzle_page())
  File "/home/tre/.local/lib/python3.8/site-packages/brozzler/cli.py", line 20
9, in brozzle_page
    browser.start(proxy=args.proxy)
  File "/home/tre/.local/lib/python3.8/site-packages/brozzler/browser.py", lin
e 352, in start
    self._wait_for(lambda: self.websock_thread.is_open, timeout=30)
  File "/home/tre/.local/lib/python3.8/site-packages/brozzler/browser.py", lin
e 323, in _wait_for
    raise BrowsingTimeout(
brozzler.browser.BrowsingTimeout: timed out after 30.0s waiting for: <function
 Browser.start.<locals>.<lambda> at 0x7f1e9549b3a0>
tre@trstreams-pc:~/Documents$
```

Restricted Mode is intended for safe code browsing. Trust this window to enable ...

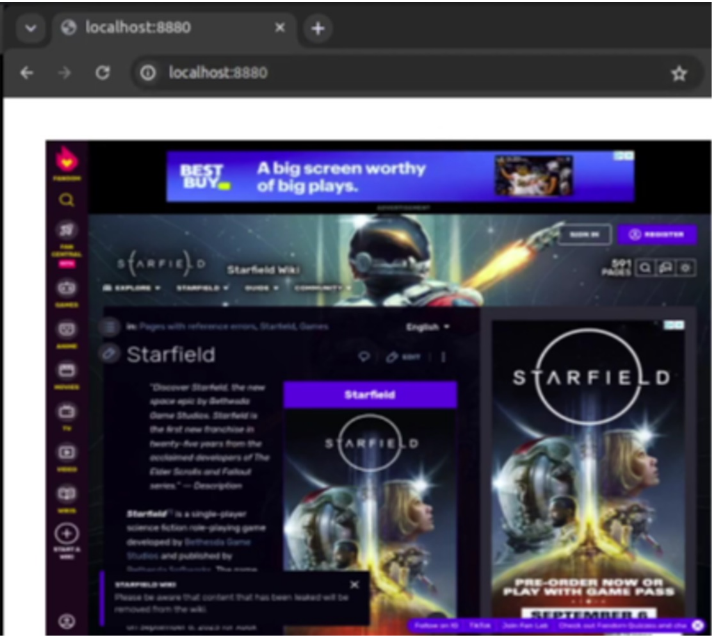
```
# Brozzle Page example
warcprox -p 8081 -d ./warcs/IGN/brozzle_page/2023_05_19
--dedup-db-file /dev/null

export
BROZZLER_EXTRA_CHROME_ARGS="--ignore-certificate-errors"

brozzle-page --chrome-exe '/usr/bin/google-chrome'
--proxy localhost:8081 'https://www.ign.com/articles/
the-last-of-us-season-1-review'


The web page does not load.
```

https://github.com/internetarchive/brozzler

Increased Support for Dynamic Content

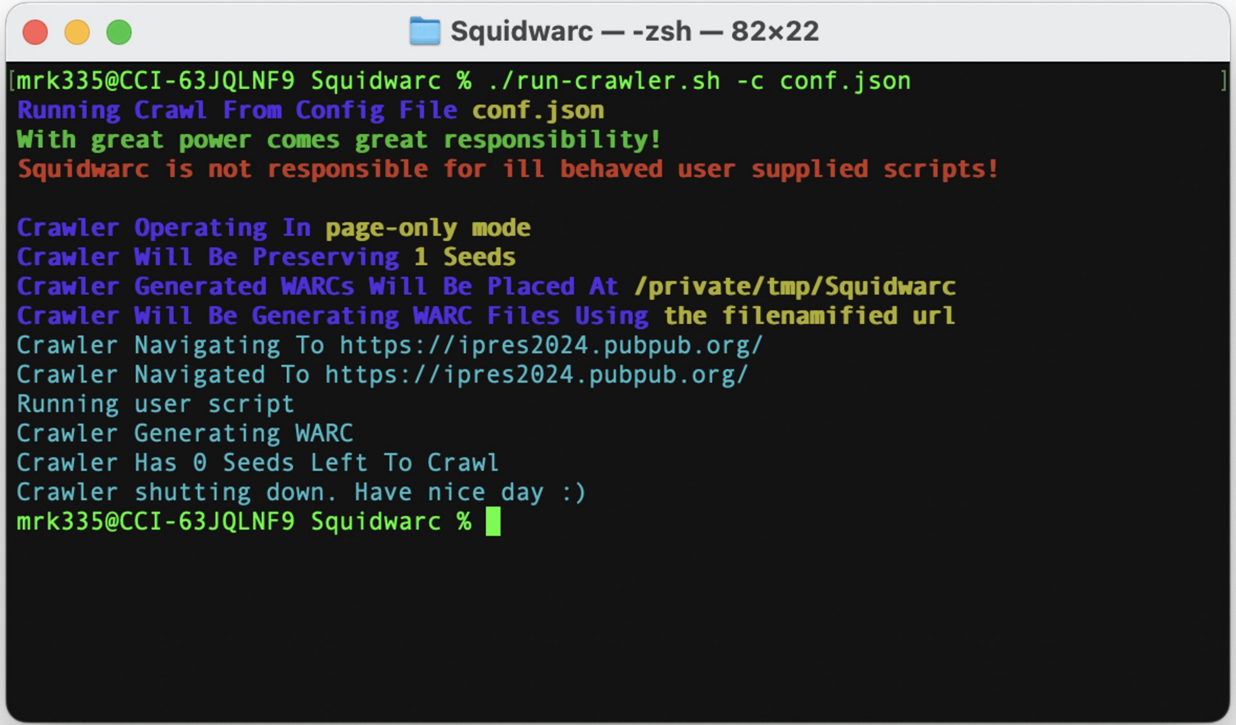


```
tre@trstreams-pc:~/Documents/livestream/collections$ docker run -p 8880:8880 -
v /home/tre/Documents/livestream/collections:/crawls -it webrecorder/browsertr
ix-crawler crawl --url https://starfield.fandom.com/wiki/Starfield --combineWA
RC --timeLimit 300 --text --allowHashUrls --behaviors autoscroll,autoplay --li
mit 1  --screencastPort=8880 --collection Starfield_Wiki
Set netIdleWait to 2 seconds
Storing state in memory
Text Extraction: Enabled

== Start:     2023-09-05 14:56:53.658
== Now:       2023-09-05 14:57:47.197 (running for 53.5 seconds)
== Progress:  0 / 1 (0.00%), errors: 0 (0.00%)
== Remaining: unknown (@ 0 pages/second)
== Sys. load: 91.8% CPU / 53.9% memory
== Workers:   1
   #0 WORK https://starfield.fandom.com/wiki/Starfield
```

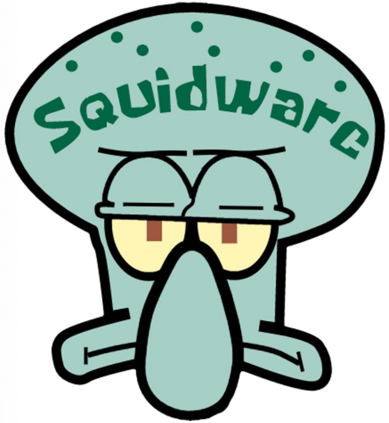


Starfield Wiki page (https://starfield.fandom.com/wiki/Starfield)

Toolbox Approach



```
mrk335@CCI-63JQLNF9 Squidwarc % ./run-crawler.sh -c conf.json
Running Crawl From Config File conf.json
With great power comes great responsibility!
Squidwarc is not responsible for ill behaved user supplied scripts!

Crawler Operating In page-only mode
Crawler Will Be Preserving 1 Seeds
Crawler Generated WARCs Will Be Placed At /private/tmp/Squidwarc
Crawler Will Be Generating WARC Files Using the filenamified url
Crawler Navigating To https://ipres2024.pubpub.org/
Crawler Navigated To https://ipres2024.pubpub.org/
Running user script
Crawler Generating WARC
Crawler Has 0 Seeds Left To Crawl
Crawler shutting down. Have nice day :)
mrk335@CCI-63JQLNF9 Squidwarc %
```

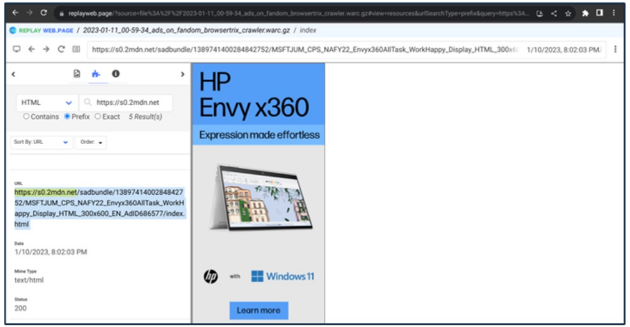


https://github.com/N0taN3rd/Squidwarc

Tools Comparison

REPLAY WEBPAGE

ReplayWeb.page's URL Search



replayweb.page/?source=file%3A%2F%2F2023-01-11_00-59-34_ads_on_fandom_browsertrix_crawler.warc.gz#view=resources&urlSearchType=prefix&query=https%3A...
REPLAY WEB.PAGE / 2023-01-11_00-59-34_ads_on_fandom_browsertrix_crawler.warc.gz / index
https://s0.2mdn.net/sadbundle/1389741400284842752/MSFTJUM_CPS_NAFY22_Envyx360AllTask_WorkHappy_Display_HTML_300x6
1/10/2023, 8:02:03 PM
HTML
https://s0.2mdn.net
○ Contains ● Prefix ○ Exact
5 Result(s)
Sort By: URL
Order:
HP
Envy x360
Expression made effortless
with Windows 11
Learn more

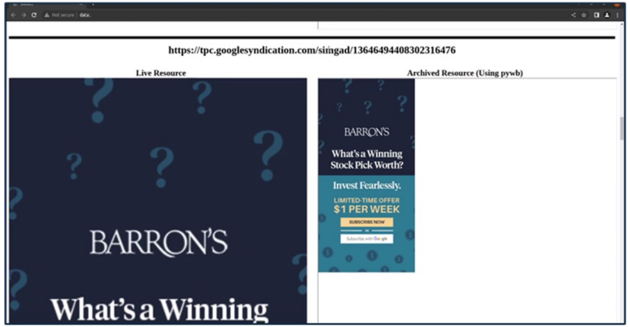
URL
https://s0.2mdn.net/sadbundle/1389741400284842752/MSFTJUM_CPS_NAFY22_Envyx360AllTask_WorkHappy_Display_HTML_300x600_EN_AdID686577/index.html

Date
1/10/2023, 8:02:03 PM

Mime Type
text/html

Status
200

Our Tool For Displaying Ads From A WARC file

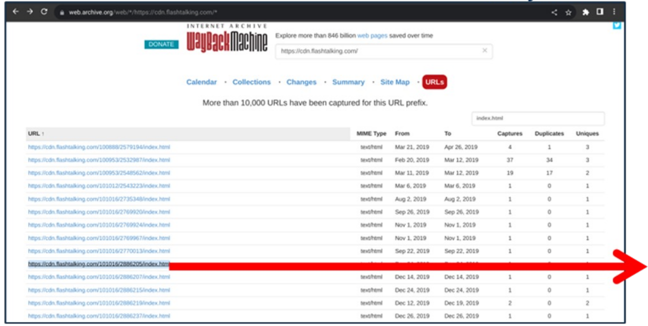


https://tpc.googlesyndication.com/simgad/13646494408302316476

Live Resource

Archived Resource (Using pywb)

Wayback Machine's URL Search

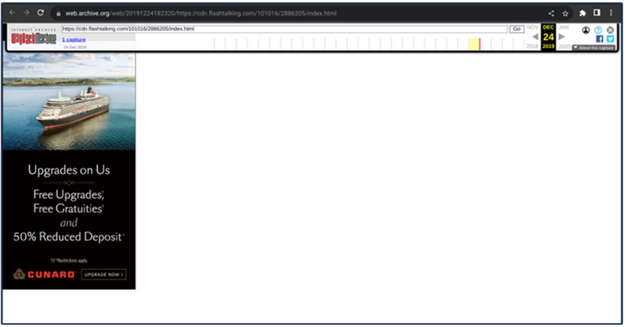


web.archive.org/web/*/https://cdn.flashtalking.com/*
INTERNET ARCHIVE
DONATE
WaybackMachine
https://cdn.flashtalking.com/
Calendar · Collections · Changes · Summary · Site Map · URLs
index.html

Explore more than 846 billion web pages saved over time

More than 10,000 URLs have been captured for this URL prefix.

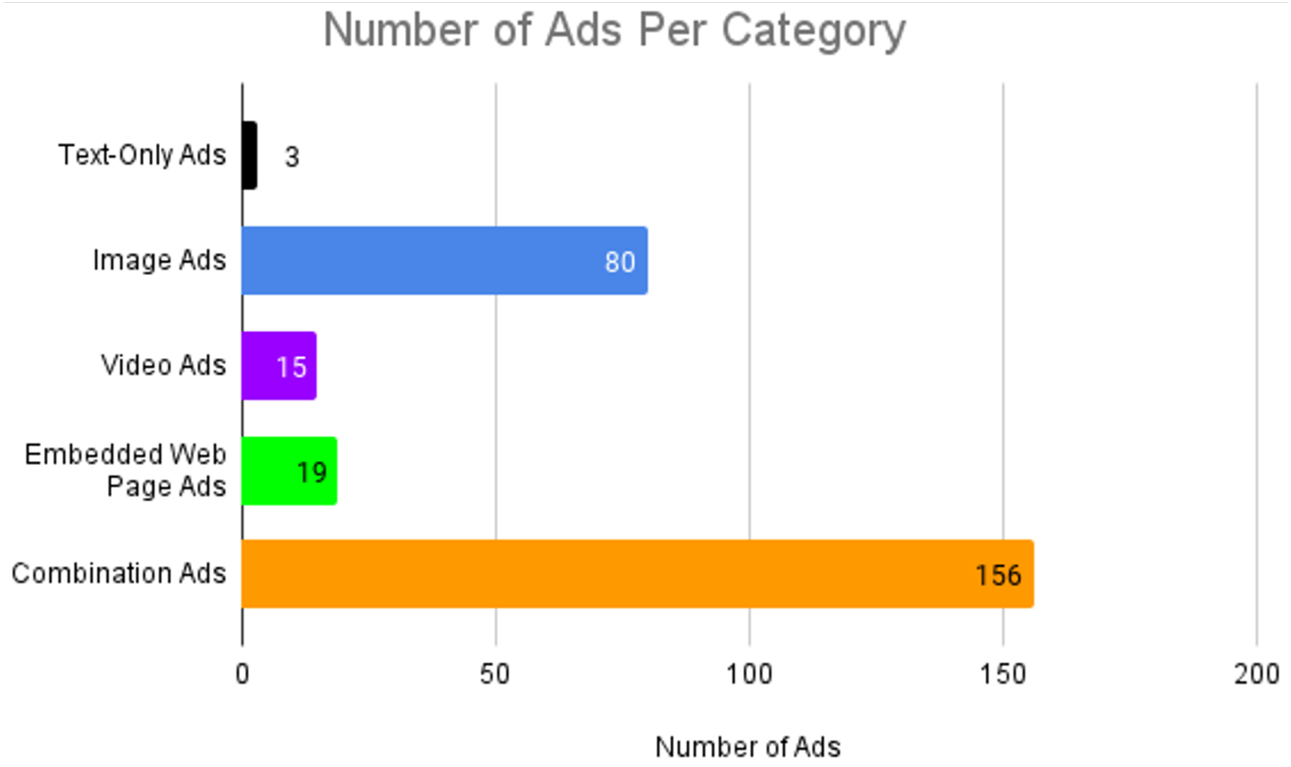
URL ↑	MIME Type	From	To	Captures	Duplicates	Uniques
https://cdn.flashtalking.com/100888/2579194/index.html	text/html	Mar 21, 2019	Apr 26, 2019	4	1	3
https://cdn.flashtalking.com/100953/2532987/index.html	text/html	Feb 20, 2019	Mar 12, 2019	37	34	3
https://cdn.flashtalking.com/100953/2548562/index.html	text/html	Mar 11, 2019	Mar 12, 2019	19	17	2
https://cdn.flashtalking.com/101012/2543223/index.html	text/html	Mar 6, 2019	Mar 6, 2019	1	0	1
https://cdn.flashtalking.com/101016/2735348/index.html	text/html	Aug 2, 2019	Aug 2, 2019	1	0	1
https://cdn.flashtalking.com/101016/2769920/index.html	text/html	Sep 26, 2019	Sep 26, 2019	1	0	1
https://cdn.flashtalking.com/101016/2769924/index.html	text/html	Nov 1, 2019	Nov 1, 2019	1	0	1
https://cdn.flashtalking.com/101016/2769967/index.html	text/html	Nov 1, 2019	Nov 1, 2019	1	0	1
https://cdn.flashtalking.com/101016/2770013/index.html	text/html	Sep 22, 2019	Sep 22, 2019	1	0	1
https://cdn.flashtalking.com/101016/2886205/index.html	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
https://cdn.flashtalking.com/101016/2886207/index.html	text/html	Dec 14, 2019	Dec 14, 2019	1	0	1
https://cdn.flashtalking.com/101016/2886215/index.html	text/html	Dec 24, 2019	Dec 24, 2019	1	0	1
https://cdn.flashtalking.com/101016/2886219/index.html	text/html	Dec 12, 2019	Dec 19, 2019	2	0	2
https://cdn.flashtalking.com/101016/2886237/index.html	text/html	Dec 26, 2019	Dec 26, 2019	1	0	1



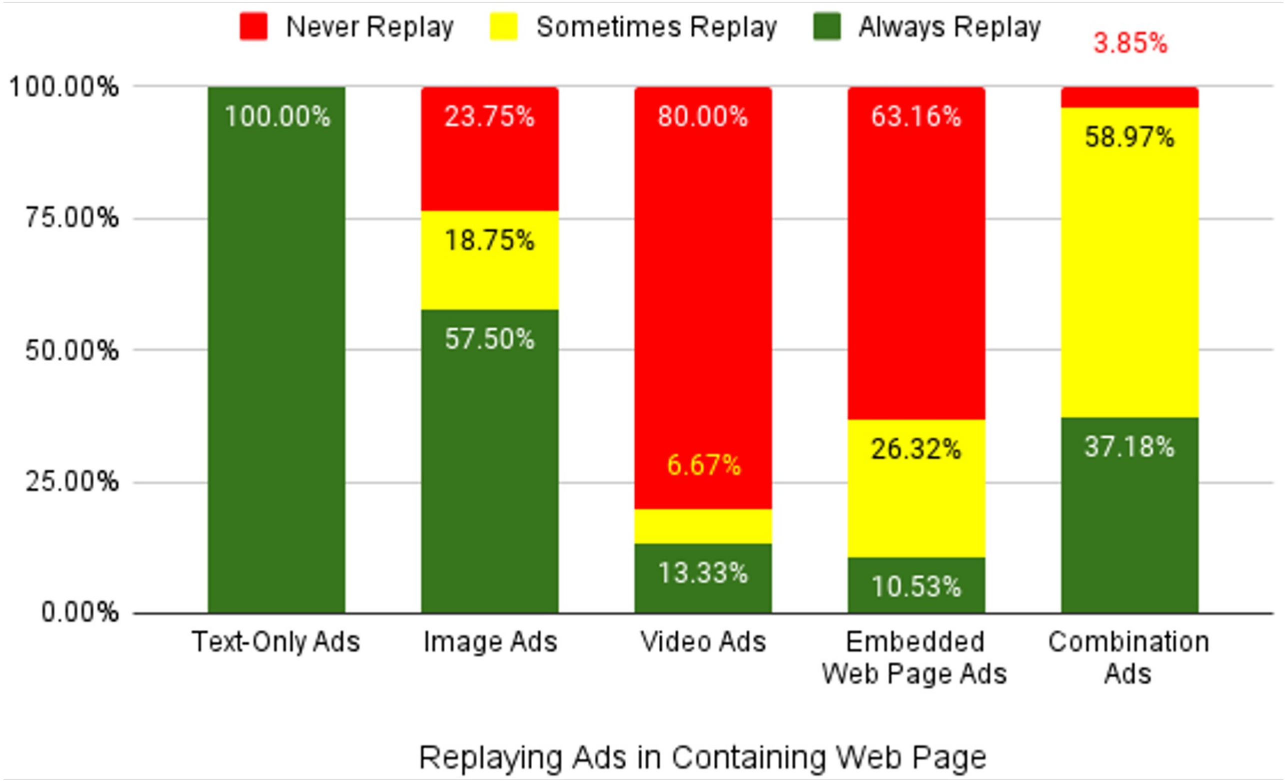
web.archive.org/web/20191224182320/https://cdn.flashtalking.com/101016/2886205/index.html
https://cdn.flashtalking.com/101016/2886205/index.html
Go
NOV DEC JAN
24
2018 2019 2020
1 capture
24 Dec 2019
Upgrades on Us
Free Upgrades,
Free Gratuities
and
50% Reduced Deposit
CUNARD
UPGRADE NOW

https://github.com/webrecorder/replayweb.page

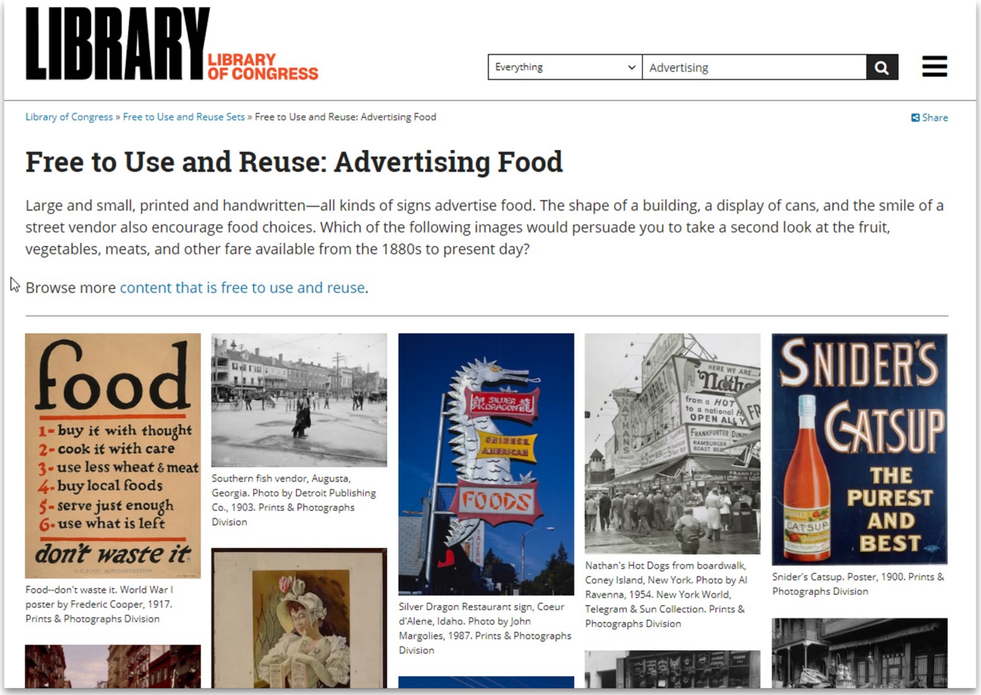
Study Results



Study Results



Pursue Partnerships and Collaboration



Library of Congress » Free to Use and Reuse Sets » Free to Use and Reuse: Advertising Food

Share

Free to Use and Reuse: Advertising Food

Large and small, printed and handwritten—all kinds of signs advertise food. The shape of a building, a display of cans, and the smile of a street vendor also encourage food choices. Which of the following images would persuade you to take a second look at the fruit, vegetables, meats, and other fare available from the 1880s to present day?

Browse more content that is free to use and reuse.

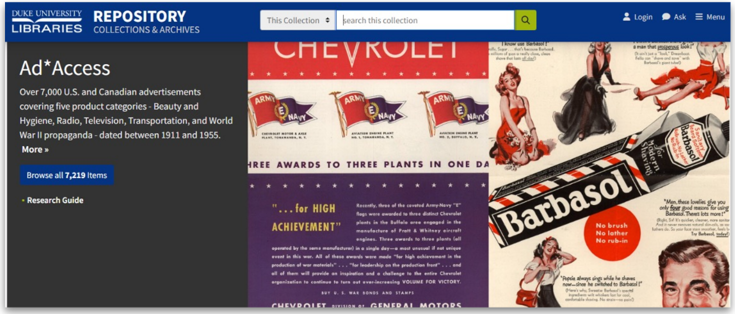
Food--don't waste it. World War I poster by Frederic Cooper, 1917. Prints & Photographs Division

Southern fish vendor, Augusta, Georgia. Photo by Detroit Publishing Co., 1903. Prints & Photographs Division

Silver Dragon Restaurant sign, Coeur d'Alene, Idaho. Photo by John Margolies, 1987. Prints & Photographs Division

Nathan's Hot Dogs from boardwalk, Coney Island, New York. Photo by Al Ravenna, 1954. New York World, Telegram & Sun Collection. Prints & Photographs Division

Snider's Catsup. Poster, 1900. Prints & Photographs Division



DUKE UNIVERSITY LIBRARIES
REPOSITORY
COLLECTIONS & ARCHIVES
This Collection
search this collection
Login
Ask
Menu
CHEVROLET
THREE AWARDS TO THREE PLANTS IN ONE DA
"...for HIGH ACHIEVEMENT"
CHEVROLET DIVISION OF GENERAL MOTORS
Barbasol
No brush
No lather
No rub-in

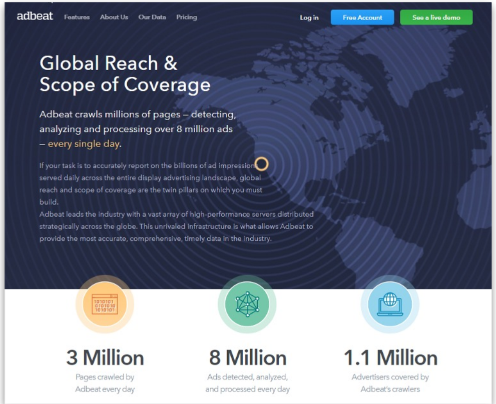
Ad*Access

Over 7,000 U.S. and Canadian advertisements covering five product categories - Beauty and Hygiene, Radio, Television, Transportation, and World War II propaganda - dated between 1911 and 1955.

More »

Browse all 7,219 Items

- Research Guide



Global Reach & Scope of Coverage

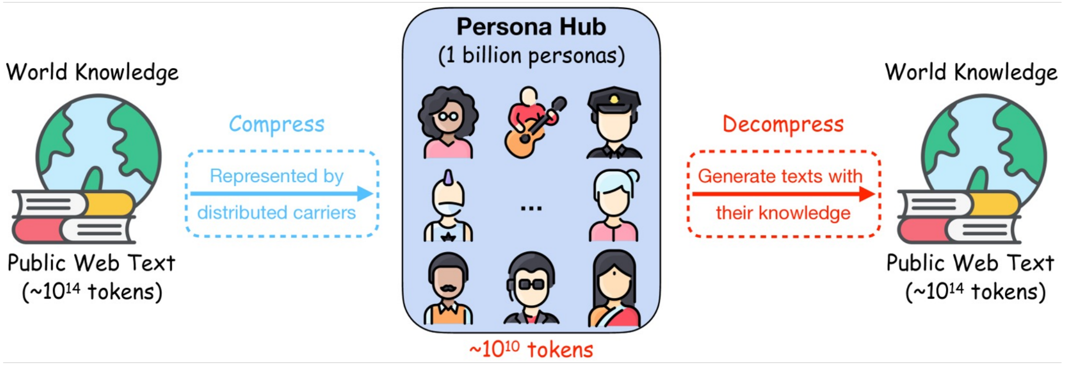
Adbeat crawls millions of pages – detecting, analyzing and processing over 8 million ads – every single day.

If your task is to accurately report on the billions of ad impressions served daily across the entire display advertising landscape, global reach and scope of coverage are the twin pillars on which you must build.

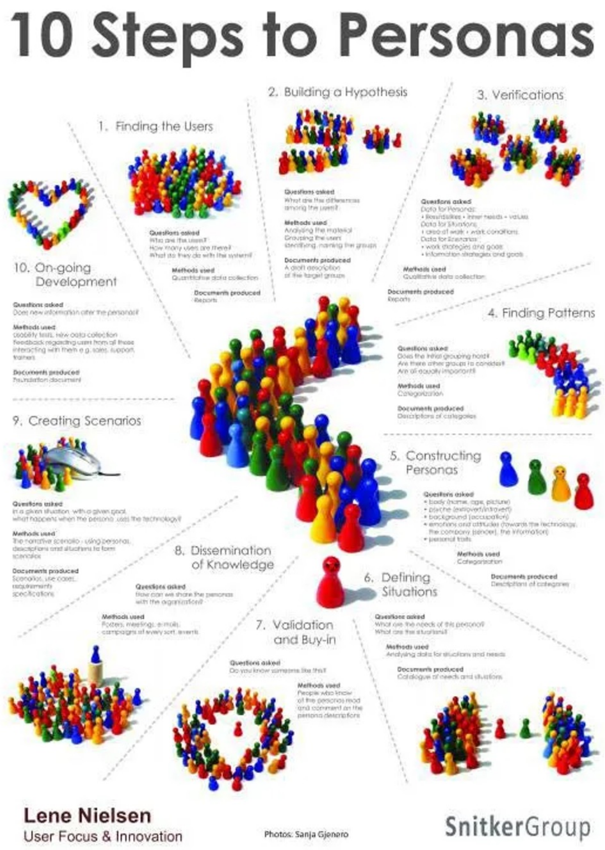
Adbeat leads the industry with a vast array of high-performance servers distributed strategically across the globe. This unrivaled infrastructure is what allows Adbeat to provide the most accurate, comprehensive, timely data in the industry.

3 Million	8 Million	1.1 Million
Pages crawled by Adbeat every day	Ads detected, analyzed, and processed every day	Advertisers covered by Adbeat's crawlers

Future Work



(Delétang et al., 2023; Ge et al., 2024), https://github.com/tencent-ailab/persona-hub


10 Steps to Personas
1. Finding the Users
Questions asked
Who are the users?
How many users are there?
What do they do with the system?
Methods used
Quantitative data collection
Documents produced
Reports
2. Building a Hypothesis
Questions asked
What are the differences among the users?
Methods used
Analysing the material
Grouping the users
Identifying, naming the groups
Documents produced
A draft description of the target groups
3. Verifications
Questions asked
Data for Personas:
• likes/dislikes • inner needs • values
Data for Situations:
• area of work • work conditions
Data for Scenarios:
• work strategies and goals
• information strategies and goals
Methods used
Qualitative data collection
Documents produced
Reports
4. Finding Patterns
Questions asked
Does the initial grouping hold?
Are there other groups to consider?
Are all equally important?
Methods used
Categorization
Documents produced
Descriptions of categories
5. Constructing Personas
Questions asked
• body (name, age, picture)
• psyche (extrovert/introvert)
• background (occupation)
• emotions and attitudes (towards the technology, the company (sender), the information)
• personal traits
Methods used
Categorisation
Documents produced
Descriptions of categories
6. Defining Situations
Questions asked
What are the needs of this persona?
What are the situations?
Methods used
Analysing data for situations and needs
Documents produced
Catalogue of needs and situations
7. Validation and Buy-in
Questions asked
Do you know someone like this?
Methods used
People who know of the personas read and comment on the persona descriptions
8. Dissemination of Knowledge
Questions asked
How can we share the personas with the organization?
Methods used
Posters, meetings, e-mails, campaigns of every sort, events
9. Creating Scenarios
Questions asked
In a given situation, with a given goal, what happens when the persona uses the technology?
Methods used
The narrative scenario - using personas, descriptions and situations to form scenarios
Documents produced
Scenarios, use cases, requirements specifications
10. On-going Development
Questions asked
Does new information alter the personas?
Methods used
Usability tests, new data collection
Feedback regarding users from all those interacting with them e.g. sales, support, trainers
Documents produced
Foundation document
Lene Nielsen
User Focus & Innovation
Photos: Sanja Gjenero
SnitkerGroup

Lene Nielsen.

In Summary

- Online ads are cultural artifacts, preservation is essentials
- Modern ads are highly personalized and dynamically loaded
- Web Ads tech is complex and the ads themselves are ephemeral
- Text/img ads are easy to capture and reply but majority use more complex techs
- Encoding personas into crawler can help to surface and preserve resources crawlers will never encounter

INSTITUTE of Museum and Library SERVICES