Web Archiving: DO IT YOURSELF

hey.

Sara Day Thomson
Digital Archivist
Centre for Research Collections
Why archive the web?
Political movements & activism

Government & corporate accountability

Research funder requirements

Reproducibility & re-use of research outputs

Personal and community digital archiving

Scholarly Citation

Influencing the world since 1583
The Web = the record of the 21st Century
Why archive web pages myself?

Image: https://digitalbevaring.dk/
Why DIY...

Web page(s) not archived by UKWA or other web archiving programme
Why DIY...

Web page(s) is archived, but is broken or incomplete
Why DIY...

Requirement for an offline copy

Image: No Wifi by Sergey Demushkin from NounProject.com
Why DIY...

Requirement for a different format, e.g. social media data
Why DIY...

Files required for project documentation, research data management, or personal digital archiving

Image: https://digitalbevaring.dk/
When to DIY...

- When citing a web resource in a publication
- As soon as a new web page goes online
- Before big changes (e.g. upgrade) or deletion of out-of-date content
- Before migrating to a new service or platform

Image: https://digitalbevaring.dk/
Web pages evolve – if you don’t begin collecting at the point of publication, the record of changes over time will be lost.
Link Rot

Web page(s) disappears or returns an error (404)

Link Drift

Content on web page or entire domain changes

Study conducted by Harvard Law School:

>70% of links in a sample of law journals don’t work

>50% of links in Supreme Court opinions don’t work

Harvard Law study: https://perma.cc/48QZ-DT6L  Influencing the world since 1583
DIY Approaches
• Check other web archives – it might already be archived
• Ask a web archive to do it for you (nominate a web page(s))
• Manually capture a web page(s) with a Webrecorder tool
• Download code and files for cheap & cheerful local copy

• Use a citation service
• Harvest machine-readable data (social media)
• Download your data directly from a platform
Check other web archives...

https://timetravel.mementoweb.org/
Check other web archives... https://ox.ac.uk etc etc
Check other web archives...

https://web.archive.org/
Ask a web archive to do it for you ...

https://web.archive.org/save/
Save a UK Website

We archive websites:

1. That are on a .uk or other UK geographic top-level domain such as .scot or .cymru
2. That are published in the UK

We do not archive:

1. Online Sound or Video platforms, in which audio-visual material is the predominant content
2. Private Intranets and Emails
3. Personal data in social networking sites or websites only available to restricted groups

Save a UK website

Fields marked with * are mandatory

Personal details you provide on this form are protected by UK Data Protection Law. Please view our Privacy Statement.

* Title of website

Please enter the title of website

* URL of the website

Please enter URL of website

Any other information

Use this space to tell us for example, if you own this website, or if it is about to go offline

Influencing the world since 1583
Manually capture a web page(s) with Webrecorder...

https://webrecorder.net/
Download code and files for cheap & cheerful local copy...

https://www.httrack.com/
Use a citation service...

Create a new Perma Link
Enter any URL to preserve it forever.

Paste your URL here.

Create Perma Link

https://perma.cc/

Disclaimer: Paid Service (legal publications)
Harvest machine-readable data (social media) ...

Type the following into your command prompt and press enter:

```
twarc search blacklivesmatter > tweets.jsonl
```

- Your collection may take some time to return all tweets. You can tell when the process has completed if your terminal has returned to the shell ($) prompt:

```
eam5hc$ twarc search blacklivesmatter > blacklivesmatter_tweets.jsonl
eam5hc$ 
```

- While the prompt is running, check to make sure your command was successful by looking inside the twarc folder. Inside you should see your tweets.jsonl file and a twarc.log file (You don’t need to open these files, just make sure they’ve been created.):

  ![Image of files in Finder](https://scholarslab.github.io/learn-twarc/)

Influencing the world since 1583
Harvest machine-readable data (social media) ...
Harvest machine-readable data (social media) ...

# TAGS
TAGS – Twitter Archiving Google Sheet

https://tags.hawksey.info/

Get TAGS

To start using TAGS select one of the versions below then follow the steps below:
Download your data directly from a platform...

[Image of a Twitter settings page with a highlighted option: "Download an archive of your data. Get insights into the type of information stored for your account."]
Download your data directly from a platform...

Download an archive of your data

Get insights into the type of information stored for your account.

Twitter data

You can request a ZIP file with an archive of your account information, account history, apps and devices, account activity, interests, and Ads data. You’ll get an in-app notification when the archive of your data is ready to download. Learn more

Twitter

Request archive

Periscope data

You can request an archive of your Periscope data on Periscope directly.

Influencing the world since 1583
Download your data directly from a platform...

- assets
- data
- Your archive.html

**Quick stats**

Here's a glance at some numbers from your archive:

- Tweets: 2,855
- Likes: 4,478
- Blocked accounts: 87
- Muted accounts: 0
- Lists: 1
- Moments: 0

---

*Hi @sdaythomson,*

Here is the information from your archive which may be most useful to you.

As this information is not all the data from your archive, please open this folder if you want to see all the data that we have included. An explanation for each piece of data can be found in the README file included in the same folder.

Date generated: March 5, 2022

Date modified: 05/03/2022 12:15

Type: File folder

Size: 41 MB

---

Home
Account
Tweets
Likes
Direct Messages
Safety
Personalization
Ads
Lists
Moments
Now for a Demo
Introducing ArchiveWeb.page – Local High-Fidelity Web Archiving

2021-01-18 • Ilya Kreynier

Introducing ArchiveWeb.page Chrome Extension

I am excited to announce the launch of ArchiveWeb.page, a brand-new high-fidelity web archiving system available as a Chrome extension from the Chrome Web Store.

The extension has been tested in latest versions of Chrome, as well as with the Edge and Brave browsers.

In classic Webrecorder style, the extension allows users to ‘record’ highly interactive websites, including social media, videos, customized content, and even local intranet content.

When the original webrecorder.io was launched nearly six years ago, the goal was to allow users to record/capture exactly what is loaded in their browser. At the time, it was not possible to do entirely with a browser extension and an outside proxy server (running on webrecorder.io) was necessary. Now, thanks to evolution of the browser technologies, this original vision of archiving entirely in your browser can finally be realized.

The ArchiveWeb.page extension turns the browser into a full web archiving system, allowing users to turn ‘recording’ mode on any tab, which will then capture/record all the elements of a page exactly as they are loaded. The archived data is then stored in the browser itself, and can be replayed/accessed even when offline. ArchiveWeb.page builds on and complements the ReplayWeb.page system, announced last year.

User Guide

To get users started with the extension right away, we’ve also launched a detailed User Guide, created by our Community Manager, Lorena Ramírez-López.

Read on below for an overview of some key features in ArchiveWeb.page.
<table>
<thead>
<tr>
<th>Crawler</th>
<th>vs</th>
<th>Webrecorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Automated</td>
<td></td>
<td>• Manual</td>
</tr>
<tr>
<td>• Capable of capturing high volume of URLs</td>
<td></td>
<td>• More effective for small number of URLs</td>
</tr>
<tr>
<td>• Typically captures minimal metadata</td>
<td></td>
<td>• Supports close qualitative analysis and more detailed metadata</td>
</tr>
<tr>
<td>• Doesn’t always capture AV or dynamic content</td>
<td></td>
<td>• Excels at capturing high fidelity copies</td>
</tr>
<tr>
<td>• Requires IT configuration and institutional infrastructure</td>
<td></td>
<td>• Easy to use for non-experts</td>
</tr>
</tbody>
</table>
https://webrecorder.net/

Webrecorder provides a suite of open source projects and tools to capture interactive websites and replay them at a later time as accurately as possible.

Latest News

Read about the ArchiveWeb.page Chrome Extension for high-fidelity web archiving.

Read about the release of the new OldWeb.today browser emulation system.

Featured Tool

ARCHIVEWEB.PAGE is a tool for high-fidelity web archiving using your browser.
API Harvesting

- Application Programming Interface – backdoor to platform
- Harvest the underlying data
  - Machine-readable (JSON, XML)
  - Rich metadata
- Only available on certain platforms

- Licenced under Terms of Service
- Supports linking datasets from different sources
- More complex personal data issues
APIs
Twitter Data

This is a test tweet @LoveforTestingT with an image.

```
{
  "created_at": "Thu Apr 30 21:53:11 +0000 2015",
  "id": 59305001022490730,
  "id_str": "59305001022490730",
  "text": "This is a test tweet @LoveforTestingT with an image, http://t.co/ZmgHvA4zdF",
  "source": "<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>",
  "truncated": false,
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": 2953882541,
    "id_str": "2953882541",
    "name": "Test Demo",
    "screen_name": "jordiee_test",
    "location": "Denver, CO",
    "url": null,
    "description": "this is a test account.",
    "protected": false,
    "verified": false,
    "followers_count": 2,
    "friends_count": 49,
    "listed_count": 0,
    "favourites_count": 0,
  }
}
```
Tweet

Elements

```json
{ "user": {
  "id": 6253282,
  "id_str": "6253282",
  "name": "Twitter API",
  "screen_name": "TwitterAPI",
  "location": "San Francisco, CA",
  "url": "https://developer.twitter.com",
  "description": "The Real Twitter API. Tweets about API changes, service issues and our Developer Platform. Don't get an answer? It's on my website.",
  "verified": true,
  "followers_count": 6129794,
  "friends_count": 12,
  "listed_count": 12899,
  "favourites_count": 31,
  "statuses_count": 3658,
  "created_at": "Wed May 23 06:01:13 +0000 2007",
}}
```
Twitter Terms & Conditions: Re-sharing

I.F.2.b.i. You may not distribute more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period, unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter.

Twitter’s Developer Policies for Researchers, Archivists, and Librarians

https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2
### #WDPD2019 search

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Link</th>
<th>Tags</th>
<th>score</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Image]</td>
<td>#WDPD2019 search</td>
<td>[Image]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[https://tags.hawksey.info/](https://tags.hawksey.info/)
Questions?

Sara.Thomson@ed.ac.uk

@idaythomson

Influencing the world since 1583