

# Archiving Websites

## Calling all researchers & packrats!

R. Scott Granneman

© 2022 R. Scott Granneman  
Last updated 2022-09-14

You are free to use this work, with certain restrictions: CC BY-SA 4.0  
For full licensing information, please see the last slide/page.

I need to save a lot of webpages

Yes, browsers support saving webpages, but it's really not very useful

Yes, you can use `wget` or `curl`, but configuring them can be ... difficult

SingleFile



Search or jump to...

Pull requests Issues Marketplace Explore



**gildas-lormeau / SingleFile** Public

Sponsor Watch Fork Star **8.8k**

Code Issues **48** Pull requests **1** Discussions Actions Wiki **6** Releases **76** 329.72 MB

master **1** branch **76** tags

<b>gildas-lormeau</b> bump version	51b70fd 2 hours ago	<b>7,114</b> commits	
<b>.github</b>	move text	...	last month
<b>_locales</b>	Merge pull request <a href="#">#985</a> from solokot/m...	...	21 days ago
<b>cli</b>	update Dockerfile	...	last month
<b>companion</b>	move companion code into "single-file-..."	...	last month
<b>lib</b>	bump version	...	2 hours ago
<b>src</b>	increase max width (see <a href="#">#998</a> )	...	2 hours ago
<b>.eslintrc.js</b>	refactor folders structure	533 B	5 months ago
<b>.gitignore</b>	format files	66 B	last month

Web Extension for Firefox/Chrome/MS Edge and CLI tool to save a faithful copy of an entire web page in a single HTML file

- javascript
- chrome-extension
- cli
- firefox
- screenshot
- chrome
- osint
- browser
- firefox-addon
- annotations
- snapshot
- selenium
- archive
- archiver
- web-extension
- add-on
- web-clipper
- puppeteer
- auto-save
- offline-reading

AGPL-3.0 license  
 **8.8k** stars  
 **108** watching  
 **646** forks

SingleFile can be installed on:

- Firefox: <https://addons.mozilla.org/firefox/addon/single-file>
- Chrome: <https://chrome.google.com/extensions/detail/mpiodijhokgodhhofbcjdecppffjipkle>
- Microsoft Edge:  
<https://microsoftedge.microsoft.com/addons/detail/efnbkdcfmcmnhlkaijjmhhjjgladedno>
- Firefox for Android Nightly by following this procedure:  
<https://blog.mozilla.org/addons/2020/09/29/expanded-extension-support-in-firefox-for-android-nightly/>



Search or jump to...

Pull requests Issues Marketplace Explore



brave / brave-browser Public

Watch Fork Starred 13.2k

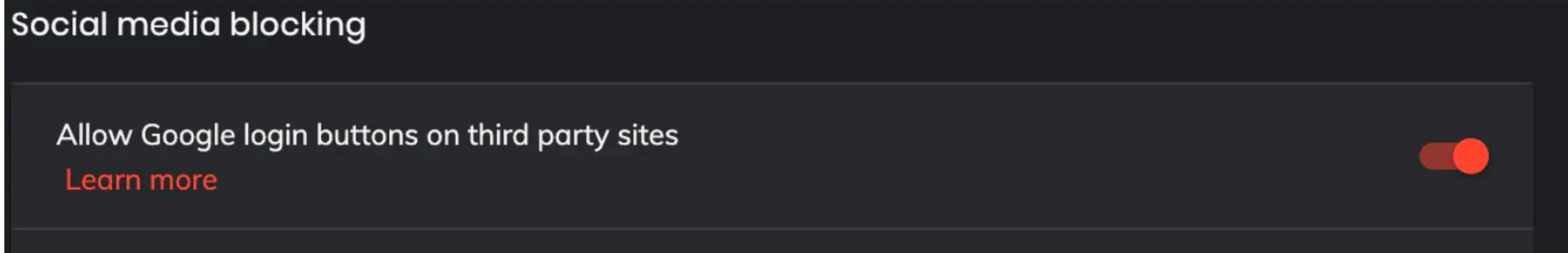
Code Issues 4.9k Pull requests 10 Actions Projects 22 Wiki 112

# Allow Google login Third Parties and Extensions

Brian Clifton edited this page 2 days ago · 6 revisions

## Social Blocking

If you navigate to `brave://settings/socialBlocking`, you will see an option to enable `Allow Google Login buttons on third party sites`.



- Pages 112
- Find a Page...
- Home
- Adding a protocol scheme to Brave
- Allow Google login Third



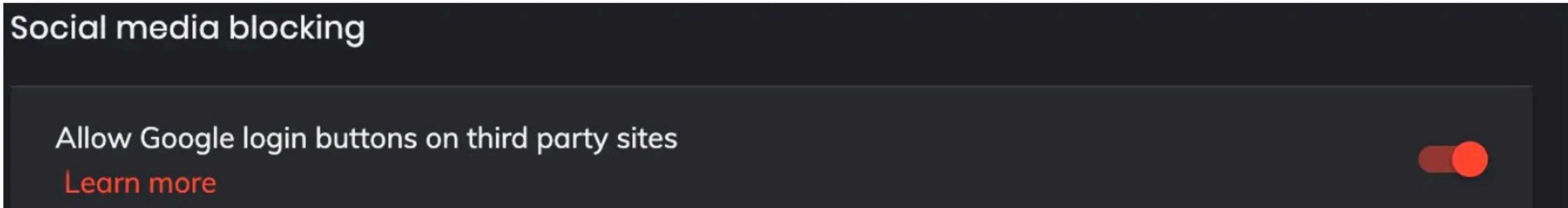


# Allow Google login Third Parties and Extensions

Brian Clifton edited this page 2 days ago · 6 revisions

## Social Blocking

If you navigate to `brave://settings/socialBlocking`, you will see an option to enable `Allow Google Login buttons on third party sites`.



Pages 112

[Home](#)

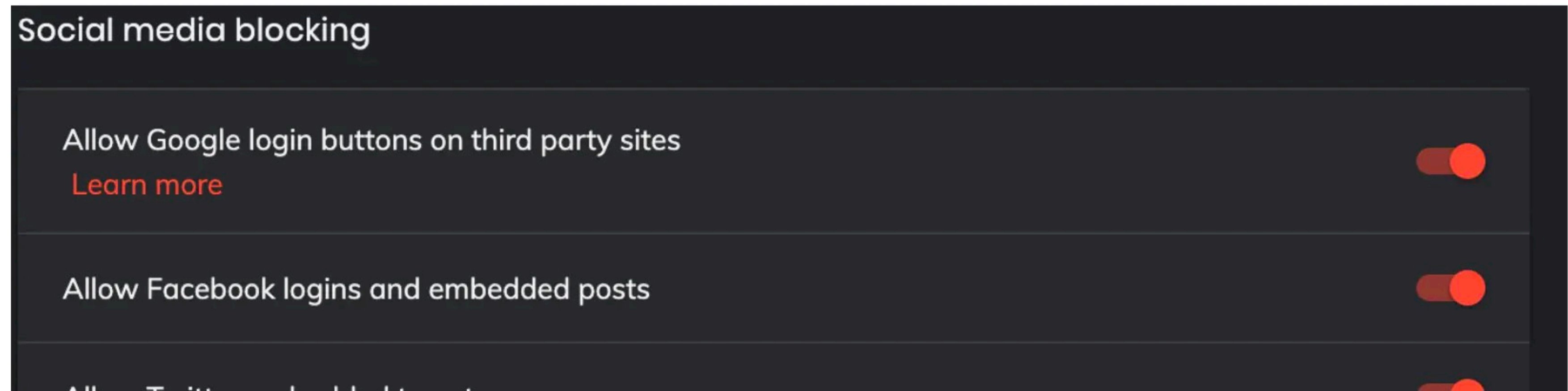
[Adding a protocol scheme to Brave](#)

# Allow Google login Third Parties and Extensions

Brian Clifton edited this page 22 hours ago · 6 revisions

## Social Blocking

If you navigate to `brave://settings/socialBlocking`, you will see an option to enable `Allow Google Login buttons on third party sites`.



Pages 112

Find a Page...

- Home
- Adding a protocol scheme to Brave
- Allow Google login Third Parties and Extensions
  - Social Blocking



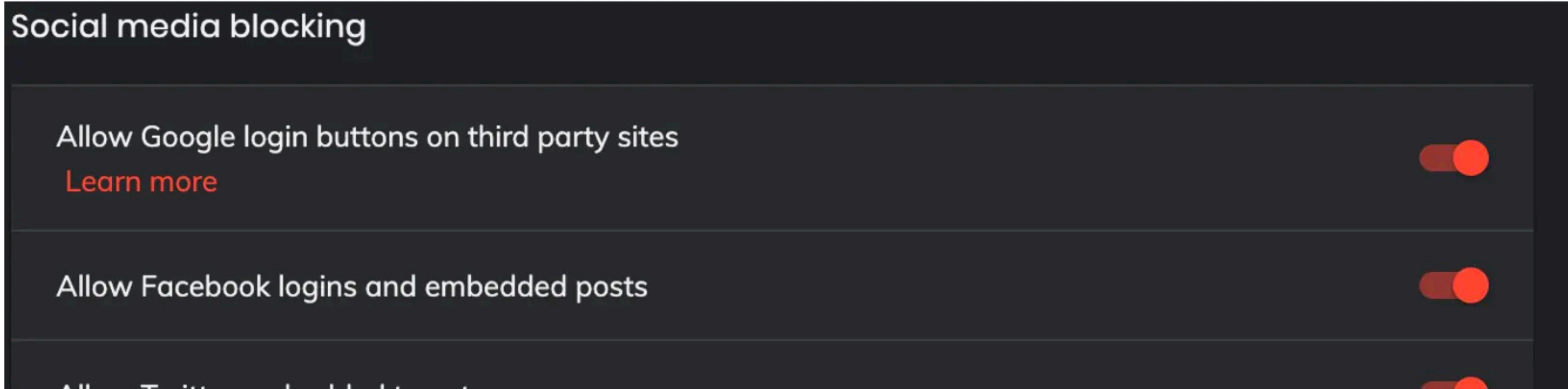
× Title: Allow Google login Third Parties and Extensions · brave/brave-browser Wiki  
URL: https://github.com/brave/brave-browser/wiki/Allow-Google-login---Third-Parties-and-Extensions  
Date: 2022-07-13

# Allow Google login Third Parties and Extensions

Brian Clifton edited this page 22 hours ago · 6 revisions

## Social Blocking

If you navigate to `brave://settings/socialBlocking`, you will see an option to enable `Allow Google Login buttons on third party sites`.



Pages 112

Find a Page...

- Home
- Adding a protocol scheme to Brave
- Allow Google login Third Parties and Extensions
  - Social Blocking

DEMO!

# Settings



SingleFile



+ **Options**

RSG



▼ User interface

add entry in the context menu of the webpage



overlay a shadow on the page during processing



display an infobar when viewing a saved page



template of the infobar content

Title: {page-title}\nURL: {url-href}\nDate: {date-iso}

open a prompt dialog to edit the infobar content



open saved pages in a new tab



auto-close the tab after the page is saved



► File name

► HTML content

► Stylesheets


► Images

► Fonts

▼ File name

template 

{page-title} - {date-iso}.html

max length 

192

bytes



open the "Save as" dialog to confirm the file name 










file name conflict resolution 

prompt for a name




## ▼ HTML content

- compress HTML content 
- remove hidden elements 
- set content security policy 
- remove frames 
- save original URLs of embedded resources 
- include the infobar in the saved page 
- save raw page 



## ▼ Stylesheets

remove unused styles 



remove stylesheets for alternative devices to screens 



compress CSS content 



move in the head element the styles found outside of it 



## ▼ Images

group duplicate images together 



save deferred images 



maximum idle time (ms) 

dispatch "scroll" event 



zoom out the page 



remove images for alternative screen resolutions 



## ▼ Fonts

remove unused fonts 



remove alternative fonts 



▼ Destination

save to filesystem

copy to clipboard

upload to GitHub

access token

user name

repository name

branch name

upload to a WebDAV server

URL

user identifier

password

upload to Google Drive

save with SingleFile Companion



▼ Network

blocked resources ⓘ

- images
- stylesheets
- fonts
- scripts
- videos
- audios

block mixed content ⓘ

"Accept" header ⓘ

- documents
- images
- stylesheets
- fonts
- scripts
- videos
- audios

set maximum size ⓘ

maximum size (MB) ⓘ

set maximum download time ⓘ

maximum download time (s) ⓘ

pass "Referer" header after a cross-origin request error ⓘ

## ▼ Annotation editor

default mode 

remove elements



apply the system theme when formatting a page 



warn if leaving page with unsaved changes 



annotate the page before saving 




open pages saved with SingleFile in the annotation editor 



## ▼ Bookmarks

save the page of a newly created bookmark 

link the new bookmark to the saved page 

ignored folders 

allowed folders 



▼ Misc.

add proof of existence 

save pages in background 

display stats in the console after processing 

---

▼ Misc.

add proof of existence



Check this option to create a worldwide proof of the existence of the page you want to save.

- **What is a proof of existence (data anchoring)?**

Data anchoring consists in building a time-stamped proof of existence for a data by linking it to a tamper resistant and time-stamped blockchain. Data anchoring implementation relies on the resilience and immutability of the Bitcoin blockchain to provide the best possible security level

- **How does this protect my data?**

The anchoring mechanism only handles data impressions. Your data remains where you calculate the fingerprints, i.e. in the browser. Their confidentiality is totally preserved.

- **The day after your backup you can get freely the proof receipt here: [gildas-lormeau.github.io/singlefile-woleet/index.html](https://gildas-lormeau.github.io/singlefile-woleet/index.html). A proof receipt will be used to verify the validity of the evidence**

More information [doc.woleet.io](https://doc.woleet.io)

save pages in background




display stats in the console after processing



ArchiveBox

# ARCHIVEBOX

 *Open source self-hosted web archiving. Takes URLs/browser history/bookmarks/Pocket/Pinboard/etc., saves HTML, JS, PDFs, media, and more...*



Search or jump to...

Pull requests Issues Marketplace Explore



ArchiveBox / ArchiveBox Public

Sponsor Watch Fork Starred 14.1k

Code Issues 139 Pull requests 15 Discussions Actions Wiki 32 Releases 27 6.33 MB

dev 14 branches 27 tags

pirate Update config.py 03eb7e5 on Jun 9 2,958 commits

.github	also install npm packages when testing ...	9 months ago
archivebox	Update config.py	3 months ago
assets	update setup.py	17 months ago
bin	fix helper install script handling of pytho...	4 months ago
brew_dist @ a43...	bump brew disk version	9 months ago
deb_dist @ f8e3...	bump deb_dist submodule	17 months ago
docker @ 236f788	add docker submodule	2 years ago
docs @ bfc5f76	bump docs version	2 years ago

Open source self-hosted web archiving. Takes URLs/browser history/bookmarks/Pocket/Pinboard/etc., saves HTML, JS, PDFs, media, and more...

archivebox.io

- python
- rss
- backups
- firefox
- pinboard
- youtube-dl
- chromium
- self-hosted
- wget
- pocket
- browser-bookmarks
- warc
- web-archiving
- wayback-machine
- digipres
- singlefile
- headless-browser
- bookmark-archiver
- internet-archiving
- archivebox

MIT license

**ArchiveBox is a powerful, self-hosted internet archiving solution to collect, save, and view sites you want to preserve offline.**

You can set it up as a [command-line tool](#), [web app](#), and [desktop app](#) (alpha), on Linux, macOS, and Windows.

**You can feed it URLs one at a time, or schedule regular imports** from browser bookmarks or history, feeds like RSS, bookmark services like Pocket/Pinboard, and more. See [input formats](#) for a full list.

**It saves snapshots of the URLs you feed it in several formats:** HTML, PDF, PNG screenshots, WARC, and more out-of-the-box, with a wide variety of content extracted and preserved automatically (article text, audio/video, git repos, etc.). See [output formats](#) for a full list.

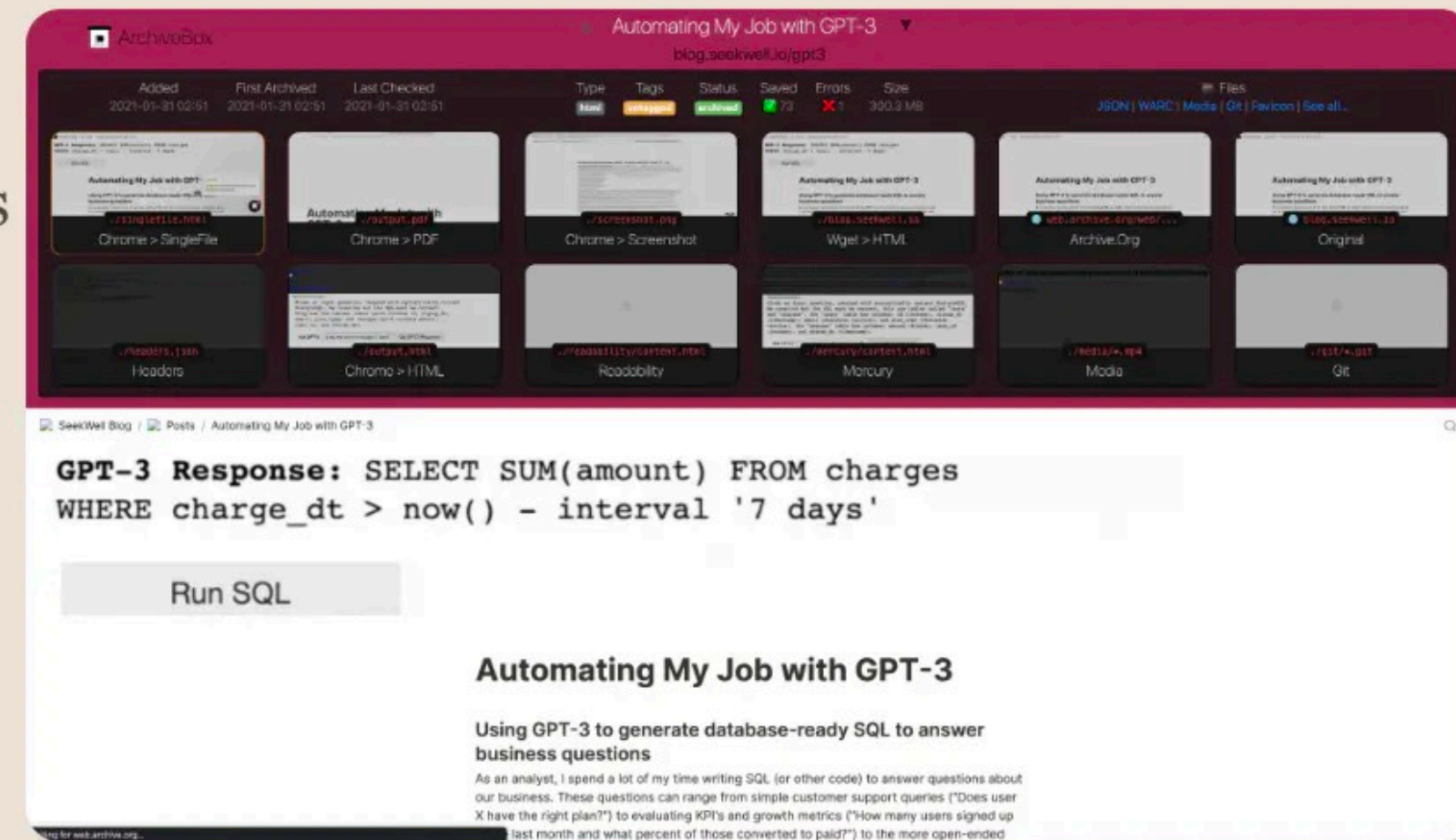


# Output Formats

Inside each Snapshot folder, ArchiveBox save these different types of extractor outputs as plain files:

`./archive/<timestamp>/*`

- › **Index:** `index.html` & `index.json` HTML and JSON index files containing metadata and details
- › **Title, Favicon, Headers** Response headers, site favicon, and parsed site title
- › **SingleFile:** `singlefile.html` HTML snapshot rendered with headless Chrome using SingleFile
- › **Wget Clone:** `example.com/page-name.html` wget clone of the site with `warc/<timestamp>.gz`
- › Chrome Headless
  - › **PDF:** `output.pdf` Printed PDF of site using headless chrome
  - › **Screenshot:** `screenshot.png` 1440x900 screenshot of site using headless chrome
  - › **DOM Dump:** `output.html` DOM Dump of the HTML after rendering using headless chrome
- › **Article Text:** `article.html/json` Article text extraction using Readability & Mercury
- › **Archive.org Permalink:** `archive.org.txt` A link to the saved site on archive.org
- › **Audio & Video:** `media/` all audio/video files + playlists, including subtitles & metadata with youtube-dl
- › **Source Code:** `git/` clone of any repository found on GitHub, Bitbucket, or GitLab links
- › *More coming soon! See the [Roadmap](#)...*






It does everything out-of-the-box by default, but you can disable or tweak [individual archive methods](#) via environment variables / config.







# Quickstart

🖥️ **Supported OSs:** Linux/BSD, macOS, Windows (Docker/WSL) 🧠 **CPUs:** amd64, x86, arm8, arm7 (raspi>=3)




## 🌟 Easy Setup

- ▶️  `docker-compose` (macOS/Linux/Windows) 📌 recommended (click to expand)
- ▶️  `docker` (macOS/Linux/Windows)
- ▶️  `bash` auto-setup script (macOS/Linux)

## 🔧 Package Manager Setup

- ▶️  `apt` (Ubuntu/Debian)
- ▶️  `brew` (macOS)
- ▶️  `pip` (macOS/Linux/Windows)
- ▶️  `pacman` /  `pkg` /  `nix` (Arch/FreeBSD/NixOS/more)

## 🏷️ Other Options

- ▶️  `docker` +  `electron` Desktop App (macOS/Linux/Windows)
- ▶️  Paid hosting solutions (cloud VPS)

## ➡️ Next Steps

- › Import URLs from some of the supported [Input Formats](#) or view the supported [Output Formats](#)...
- › Tweak your UI or archiving behavior [Configuration](#) or read about some of the [Caveats](#) and troubleshooting steps...
- › Read about the [Dependencies](#) used for archiving, the [Upgrading Process](#), or the [Archive Layout](#) on disk...
- › Or check out our full [Documentation](#) or [Community Wiki](#)...

# CLI

```
# archivebox [subcommand] [--args]
# docker-compose run archivebox [subcommand] [--args]
# docker run -v $PWD:/data -it [subcommand] [--args]

archivebox init --setup      # safe to run init multiple times (also how you update versions)
archivebox --version
archivebox help
```

# Start Web UI

```
archivebox manage createsuperuser # set an admin password
archivebox server 0.0.0.0:8000 # open http://127.0.0.1:8000 to view it

# you can also configure whether or not login is required for most features
archivebox config --set PUBLIC_INDEX=False
archivebox config --set PUBLIC_SNAPSHOTS=False
archivebox config --set PUBLIC_ADD_VIEW=False
```

Select snapshot to change | In: x Archived Sites x +

Not Secure | 0.0.0.0:8000/add/ Incognito

ArchiveBox: Add Links | Admin | Docs  
Index

# Add new URLs to your archive

URLs (one per line):

```
https://example.com
https://example.com/1234
some text [with](https://example.com/urls/in/it) works too|
```

Archive depth:

depth = 0 (archive just these URLs)

depth = 1 (archive these URLs and all URLs one hop away)

Archive methods:

title
favicon
wget
singlefile

**Add URLs and archive +**

Bookmark this link to quickly add to your archive: [Add to ArchiveBox](#)

Archive created using [ArchiveBox](#) version [v0.5.3](#).

Q [Title, URL, tags, timestamp, or content...]

Search

1-50 of 13056 total (Page 1 of 262)



BOOKMARKED	SNAPSHOT (13056)	FILES	ORIGINAL URL
2022-09-12 7:19PM	abdu 🎁 (@abdu8ya) nitter	6	<a href="https://nitter.net/abdu8ya">https://nitter.net/abdu8ya</a>
2022-09-12 7:19PM	abdu 🎁 (@abdu8ya): "Guys, I need to come clean. I'm not a YIMBY because I want more housing. I'm a YIMBY because I really love ...	4	<a href="https://nitter.net/abdu8ya/status/1556817069131300864#...">https://nitter.net/abdu8ya/status/1556817069131300864#...</a>
2022-09-12 7:12PM	MVP.css - Minimalist stylesheet for HTML elements	6	<a href="https://andybrewer.github.io/mvp">https://andybrewer.github.io/mvp</a>
2022-09-12 7:12PM	Cancer breakthrough is a 'wake-up' call on danger of air pollution   Cancer research   The Guardian	6	<a href="https://theguardian.com/science/2022/sep/10/cancer-bre...">https://theguardian.com/science/2022/sep/10/cancer-bre...</a>
2022-09-10 7:14PM	As Burning Man Goes Virtual, Organizers Try To Capture The Communal Aspect : NPR	5	<a href="https://npr.org/2020/09/03/908767529/as-burning-man-go...">https://npr.org/2020/09/03/908767529/as-burning-man-go...</a>
2022-09-08 7:12PM	Accelerate Python code 100x by import taichi as ti   Taichi Docs	5	<a href="https://docs.taichi-lang.org/blog/accelerate-python-code-1...">https://docs.taichi-lang.org/blog/accelerate-python-code-1...</a>
2022-09-08 7:12PM	Burning Man 2022 and the embezzling of Empyrean   by Graham Berry   Aug, 2022   Medium	5	<a href="https://medium.com/@grahamberry/burning-man-2022-an...">https://medium.com/@grahamberry/burning-man-2022-an...</a>
2022-09-07 7:23PM	Workhorse Jersey - Preorder - Mosko Moto	6	<a href="https://moskomoto.com/products/workhorse-jersey">https://moskomoto.com/products/workhorse-jersey</a>
2022-09-07 7:23PM	Juul Settles Multistate Youth Vaping Inquiry for \$438.5 Million - The New York Times	4	<a href="https://nytimes.com/2022/09/06/health/juul-settlement-va...">https://nytimes.com/2022/09/06/health/juul-settlement-va...</a>
2022-08-27 7:11PM	The Silence of Risk Management Victory - by Stephanie Losi	6	<a href="https://riskmusings.substack.com/p/the-silence-of-risk-ma...">https://riskmusings.substack.com/p/the-silence-of-risk-ma...</a>
2022-08-25 7:19PM	ani betts (@anaisbetts) nitter	5	<a href="https://nitter.net/anaisbetts">https://nitter.net/anaisbetts</a>
2022-08-25 7:19PM	ani betts (@anaisbetts): "We built all of GitHub, with zero initial funding, with absolutely no grind culture of any kind, ever..."	5	<a href="https://nitter.net/anaisbetts/status/156247755331486105...">https://nitter.net/anaisbetts/status/156247755331486105...</a>
2022-08-25 7:19PM	ani betts (@anaisbetts): "We built all of GitHub, with zero initial funding, with absolutely no grind culture of any kind, ever..."	5	<a href="https://nitter.net/i/status/1562477553314861058">https://nitter.net/i/status/1562477553314861058</a>
2022-08-25 7:19PM	Oven (Bun) is hiring engineers (@oven_sh) nitter	5	<a href="https://nitter.net/oven_sh">https://nitter.net/oven_sh</a>

This thing is amazing! We should all be using it immediately!

## Storage Requirements

Because ArchiveBox is designed to ingest a firehose of browser history and bookmark feeds to a local disk, it can be much more disk-space intensive than a centralized service like the Internet Archive or Archive.today. **ArchiveBox can use anywhere from ~1gb per 1000 articles, to ~50gb per 1000 articles**, mostly dependent on whether you're saving audio & video using `SAVE_MEDIA=True` and whether you lower `MEDIA_MAX_SIZE=750mb`.

Disk usage can be reduced by using a compressed/deduplicated filesystem like ZFS/BTRFS, or by turning off extractors methods you don't need. **Don't store large collections on older filesystems like EXT3/FAT** as they may not be able to handle more than 50k directory entries in the `archive/` folder. **Try to keep the `index.sqlite3` file on local drive (not a network mount)** or SSD for maximum performance, however the `archive/` folder can be on a network mount or spinning HDD.



Thank you!

[scott@granneman.com](mailto:scott@granneman.com)

[www.granneman.com](http://www.granneman.com)

[@scottgranneman](https://twitter.com/scottgranneman)

# Archiving Websites

## Calling all researchers & packrats!

R. Scott Granneman

© 2022 R. Scott Granneman  
Last updated 2022-09-14

You are free to use this work, with certain restrictions: CC BY-SA 4.0  
For full licensing information, please see the last slide/page.

# Changelog

2022-09-14 1.0: Created & finished slides

# Licensing of this work

This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>.

You are free to:

- » *Share* — copy and redistribute the material in any medium or format
- » *Adapt* — remix, transform, and build upon the material for any purpose, even commercially

Under the following terms:

*Attribution.* You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. Give credit to:

**Scott Granneman • [www.granneman.com](http://www.granneman.com) • [scott@granneman.com](mailto:scott@granneman.com)**

*Share Alike.* If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

*No additional restrictions.* You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Questions? Email [scott@granneman.com](mailto:scott@granneman.com)