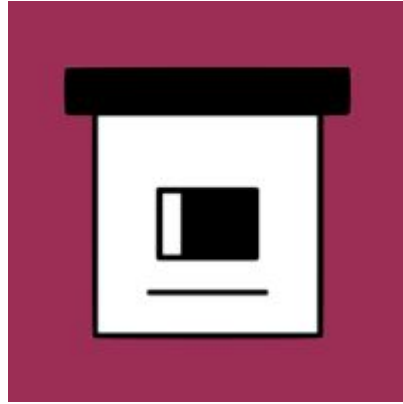


Web archiving with ArchiveBox



Tom Ryder

tom@sanctum.geek.nz
<https://sanctum.geek.nz/>

Quoth the server:

404

The fickle web—1/4

- Do you still keep a lot of bookmarks in your browser?
- How many of them still work?
- How many of them redirect somewhere unhelpful?
- How many of them yield a 404?
- How many of them time out?
- How many don't even connect?



The fickle web—2/4

- How many of you have lost something special that was published on the internet, now gone forever?
 - A work of art?
 - Photos of a loved one?
 - Posts from friends long gone?



The fickle web—3/4

- Many parties are involved in keeping content online:
 - Author...
 - Editor...
 - Advertiser...
 - Publisher...
 - Web host...
 - Sysadmin...
 - Domain registrar...
- If any of them don't do their jobs, things disappear.



The fickle web—4/4

- [Link rot](#) is the natural decay of links into no longer retrieving their content
- It's a natural consequence of how the web works...
- ...but it's getting pretty bad:
 - “A 2013 study found that 49% of links in U.S. Supreme court opinions are dead.”
—[The Atlantic](#)



Remediation—1/2

- Companies and publishers investing in diligent content curation, for the long-term public good, and the health of the open internet.





Remediation—2/2

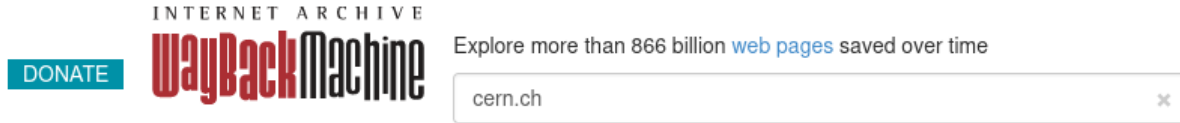
- ~~Companies and publishers investing in diligent content curation, for the long-term public good, and the health of the open internet.~~
- Archive Team
- [archive.today](#)
- Internet Archive
 - In particular, the [Wayback Machine](#).



The Wayback Machine—1/3

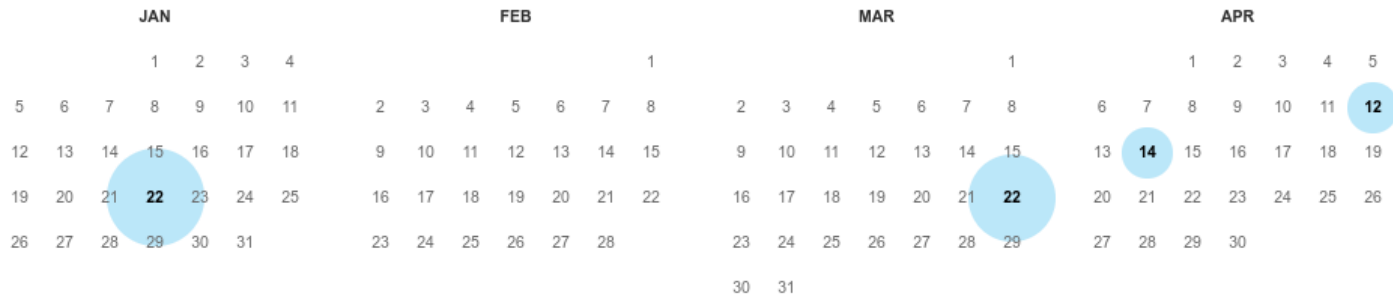
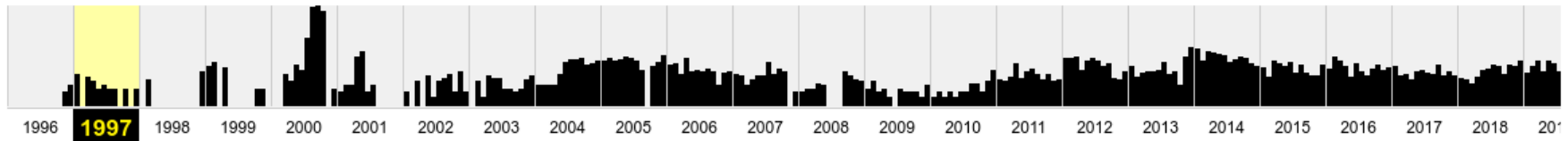
- **Billions** of historical web page snapshots—since **1996** (!)
- Great for tracking changes over time
- Still crawls the web constantly
- Users can request snapshots of given pages

The Wayback Machine—2/3



[Calendar](#) · [Collections](#) · [Changes](#) · [Summary](#) · [Site Map](#) · [URLs](#)

Saved **13,096 times** between [November 15, 1996](#) and [February 10, 2024](#).



The Wayback Machine—3/3

http://www.cern.ch/ Go JAN MAR MAY
13,096 captures
15 Nov 1996 - 10 Feb 2024
1996 1997 1998 About this capture



European Laboratory for Particle Physics

[Lab](#) - [News](#) - [Activities](#) - [Physics](#) - [Other Subjects](#) - [Index](#) - [Search](#) - [Shrink](#) - Expand



Welcome to the European Laboratory for [Particle Physics](#), located near [Geneva](#) in [Switzerland](#) and [France](#). CERN is the birthplace of the [World-Wide Web](#).

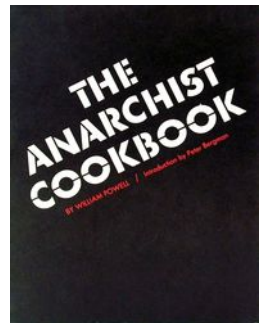
The Golden Rule

- If you find it on the internet...
- ...and you like it...
- ...save a **local copy** of it.
 - Not a bookmark!
 - Not to cloud storage!
 - A **copy**, to your **computer**, with **backups**.



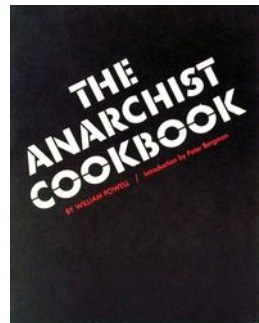
The Archivist Cookbook—1/3

- The public web archive systems and tools are great, but they still have a few problems:
 - They can't save **everything**—the web is enormous.
 - They're **centralised**—you're still trusting someone else to look after your content.
 - They're **public**—you can't keep *private* snapshots of web pages, or snapshots of *private* sites (e.g. intranets).



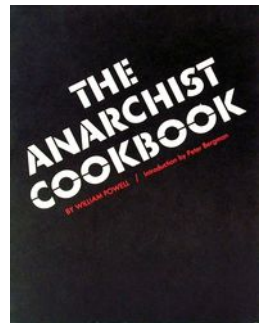
The Archivist Cookbook—2/3

- In the browser: **Right click, Save Page As...!**
 - Hint: you almost always want “Web page, complete”
- Not always that simple.
- Multiple methods of archiving are required.
- Dedicated tools help here:
 - [yt-dlp](#) for videos
 - [gallery-dl](#) for image sets



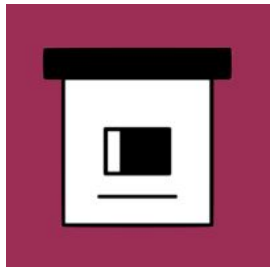
The Archivist Cookbook—3/3

- What if you could have your own personal Wayback-Machine-style web archive?
- What if each time you find a page you want to save, you could click a button, and it automatically saves it to a local database, for your use only?



Enter ArchiveBox—1/3

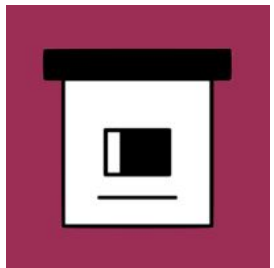
- **Free and open-source software**
 - Python ([Django](#)), JavaScript ([Node.js](#))
- Stores content in ordinary files and folders
- Real-time or scheduled snapshots
- Standard, long-term formats (HTML, PDF, PNG, WARC...)
- Managed by web frontend or CLI
- Runs in [Docker](#) (if you're into that...)



Enter ArchiveBox—2/3

Archives web pages via several different methods automatically, including:

- Single file HTML (with scripts, styles, images...)
- Full-page screenshot
- Printable PDF
- WARC archive format
- Text content
- Videos (yt-dlp)
- Git repositories



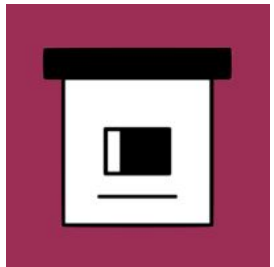
Enter ArchiveBox—3/3

BOOKMARKED	SNAPSHOT (482)	FILES	ORIGINAL URL
2024-01-26 10:57AM	Installation · roundcube/roundcubemail Wiki · GitHub	12	https://github.com/roundcube/roundcubem...
2024-01-26 10:55AM	Garbage collector does not work · Issue #6323 · roundcube/roundcubemail · GitHub	12	https://github.com/roundcube/roundcubem...
2024-01-23 10:50PM	10.4.7 - Cerb	11	https://cerb.ai/releases/10.4.7/
2024-01-23 10:46PM	Release Roundcube Webmail 1.6.6 · roundcube/roundcubemail · GitHub	12	https://github.com/roundcube/roundcubem...
2024-01-23 10:46PM	Update 1.6.6 released	11	https://roundcube.net/news/2024/01/20/up...
2024-01-19 1:57PM	Upgrading Phabricator	11	https://secure.phabricator.com/book/phabri...
2024-01-15 2:06PM	Direct System Log Messages to a Remote Destination Junos OS Juniper Networks	8	https://www.juniper.net/documentation/us/e...
2024-01-12 9:59AM	SMTP Smuggling	8	https://www.postfix.org/smtp-smuggling.ht...
2024-01-12 9:53AM	NZ Defence Force disables internet access for staff after unspecified issue emerges - NZ Herald	8	https://www.nzherald.co.nz/nz/nz-defence-f...
2024-01-10 3:36PM	Error with PHP 8.4 WordPress.org	8	https://wordpress.org/support/topic/error-wi...
2024-01-05 2:42PM	PHP: Description of core php.ini directives	8	https://www.php.net/manual/en/ini.core.php...

Installation

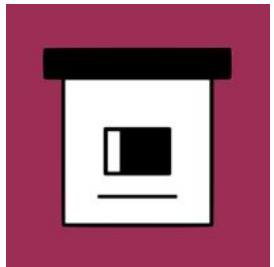
- Multiple methods, including a pre-made Docker image.
- I favor the **PyPI package** in a **venv**:

```
$ python3 -m venv archivebox  
$ . archivebox/bin/activate  
$ pip install archivebox
```



Create a new archive

```
$ mkdir ~/archive  
$ cd ~/archive  
$ archivebox init
```



Setup—1/2

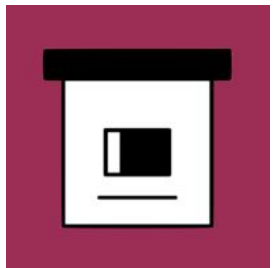
1) Install Chromium, curl, git, Node.js, npm, ripgrep, wget, and yt-dlp (keep that one up-to-date):

– **Debian/Ubuntu:**

```
$ sudo apt install chromium curl git nodejs  
npm ripgrep wget  
$ pip install yt-dlp
```

2) Run `archivebox setup`.

3) Provide an admin **username** and **password**.



Setup—2/2

```
[√] Set up ArchiveBox and its dependencies successfully.
0.7.2
ArchiveBox v0.7.2 BUILD_TIME=2024-02-13 08:12:08 1707811928
IN_DOCKER=False IN_QEMU=False ARCH=x86_64 OS=Linux PLATFORM=Linux-6.6.13-amd64-x86_64-with-glibc2.37 PYTHON=Cpython
FS_ATOMIC=True FS_REMOTE=False FS_USER=1000:1000 FS_PERMS=644
DEBUG=False IS_TTY=True TZ=UTC SEARCH_BACKEND=ripgrep LDAP=False

[i] Dependency versions:
✓ PYTHON_BINARY          v3.11.7      valid      /usr/bin/python3.11
✓ SQLITE_BINARY          v2.6.0       valid      /usr/lib/python3.11/sqlite3/dbapi2.py
✓ DJANGO_BINARY           v3.1.14     valid      .box/lib/python3.11/site-packages/django/__init__.py
✓ ARCHIVEBOX_BINARY       v0.7.2       valid      .box/bin/archivebox

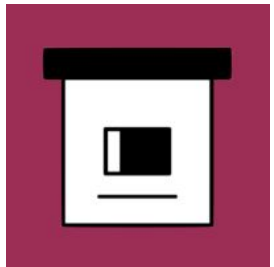
✓ CURL_BINARY             v8.5.0       valid      /usr/bin/curl
✓ WGET_BINARY              v1.21.4     valid      /usr/bin/wget
✓ NODE_BINARY              v18.19.0    valid      /usr/bin/node
✓ SINGLEFILE_BINARY        v1.1.50     valid      ./node_modules/single-file-cli/single-file
✓ READABILITY_BINARY       v0.0.11     valid      ./node_modules/readability-extractor/readability-extractor
✓ MERCURY_BINARY           v1.0.0       valid      ./node_modules/@postlight/parser/cli.js
✓ GIT_BINARY               v2.43.0     valid      /usr/bin/git
✓ YOUTUBEDL_BINARY         v2023.12.30 valid      .box/bin/yt-dlp
✓ CHROME_BINARY            v121.0.6167.160 valid      /usr/bin/chromium
✓ RIPGREP_BINARY           v14.1.0     valid      /usr/bin/rg
```

Run the webserver


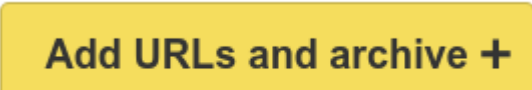
1) Run `archivebox` server.

2) Browse to `http://127.0.0.1:8000/`

- You can put a proper hostname and HTTPS on it too, using a reverse proxy: [Apache HTTPD](#), or [Caddy](#), or [Nginx](#)...
- ...but maybe that's just me.
- You'll want to **automate** starting the server.



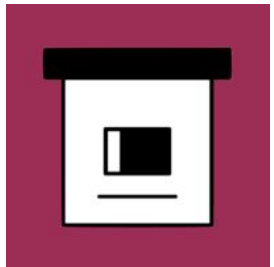
Create a new snapshot

- Click **ADD:** 
- Provide the username and password you gave in the setup step.
- Provide at least one URL, and: 

Add new URLs to your archive

URLs (one per line):











<https://www.plug.org.nz/>



Voilà!—1/2

ArchiveBox ADD + SNAPSHOTS | TAGS | LOG DOCS | PUBLIC | ADM

Q Search Tags + - ↓ Title Pull Re-Snapshot Reset Delete 0 of 1 selected

<input type="checkbox"/>	ADDED	TITLE	FILES SAVED	SIZE	ORIGINAL URL
<input type="checkbox"/>	2024-02-13 9:32PM	PLUG – Palmerston North Linux Users Group	         	11.0 MB	https://www.plug.org.nz/

1 snapshot



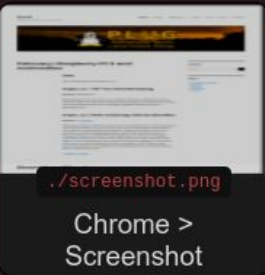




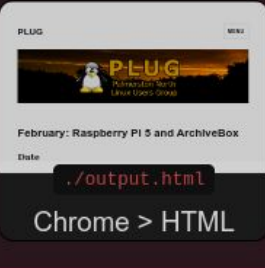




Voilà!—2/2

ArchiveBox **PLUG – Palmerston North Linux Users Group**
https://www.plug.org.nz/

Added: 2024-02-13 08:32 | First Archived: 2024-02-13 08:32 | Last Checked: 2024-02-13 08:32

Tags (html) | Status: **archived** | Saved: ✓ 9 | Errors: × 3 | Size: 11.0 MB | Snapshot: [1707813136.447049] e48425ff-c449-4b96-a110...

[JSON](#) | [WARC](#) | [Media](#) | [Git](#) | [Actions](#) | [Admin](#) | [See all files...](#)

 <p>./singlefile.html</p> <p>Chrome > SingleFile</p>	 <p>./output.pdf</p> <p>Chrome > PDF</p>	 <p>./screenshot.png</p> <p>Chrome > Screenshot</p>	 <p>./www.plug.org.nz</p> <p>Wget > HTML</p>	 <p>web.archive.org/web/...</p> <p>Archive.Org</p>	 <p>www.plug.org.nz</p> <p>Original</p>
 <p>./headers.json</p> <p>Headers</p>	 <p>./output.html</p> <p>Chrome > HTML</p>	 <p>./readability/content.html</p> <p>Readability</p>	 <p>./mercury/content.html</p> <p>Mercury</p>	 <p>./media/*.mp4</p> <p>Media</p>	 <p>./git/*.git</p> <p>Git</p>

Snapshots en-masse

- You can also add new snapshots from the command line:

```
$ archivebox add https://www.plug.org.nz/
```

- Great for archiving all your browser bookmarks in one hit!

```
$ archivebox add < firefox-bookmarks.html
```

Tips and tricks

- Check the different methods to see which ones worked best for your site.
 - The “singlefile” method is my usual go-to...when it works, which it doesn't always.
- Bookmark the “Add” page of your instance.
- Keep backups of your archive.
- Disable submitting snapshots to Internet Archive, if you have privacy concerns. This feature defaults to being on, which I think is a mistake.
`SAVE_ARCHIVE_DOT_ORG=False`
- Don't archive JavaScript on dodgy sites. Security issues still apply...

Questions?

- Website
- Source code
- Background & Motivation
- Quickstart

Email: tom@sanctum.geek.nz

Website: <https://sanctum.geek.nz/>

Fediverse: [@tejr@mastodon.sdf.org](https://mstdn.social/@tejr)

