



# Archiving the Internet Before it All Rots Away

**Nick Sweeting (@theSquashSH)**

**PyCon Colombia 2020**



Monadical Inc.

# Nick Sweeting

[twitter.com/theSquashSH](https://twitter.com/theSquashSH) [github.com/pirate](https://github.com/pirate)

Co-Founder @ Monadical.com



**We're a software development consultancy!**

- Python & JS full-stack work
- Lots of open-source + interesting projects
- Fully remote & flexible hours

Disclaimer: I am not a digital preservation professional,  
I don't work for any archiving organizations. I just think it's neat.



## 2 min: Self intro

- name, company
- founded in Colombia
- poker -> consulting, fully remote in MTL and NYC now

Why am I into internet  
archiving?

Well... where I grew up...

- my connection was 1.5mbps on a good day, with ~200ms ping
- the internet was heavily censored (often unpredictably)
- Google something sketchy and your house goes offline for 30min

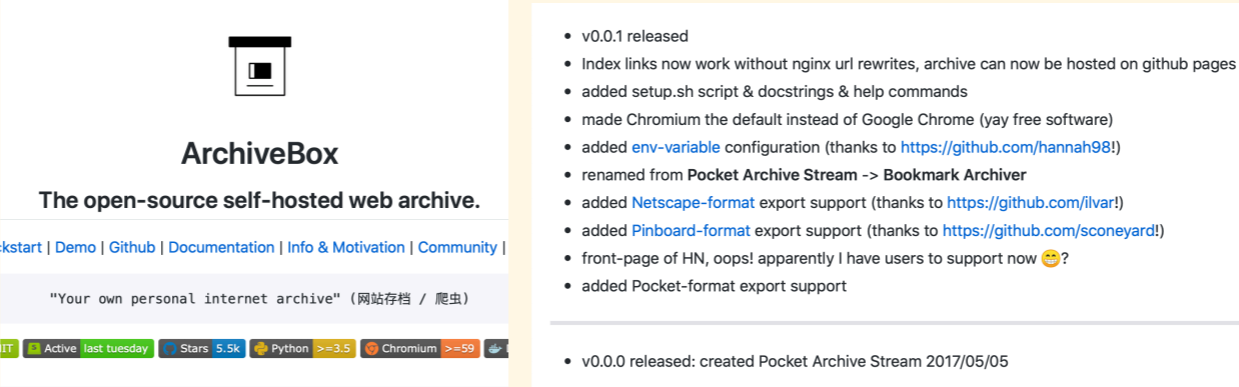
Having reliable access to content  
is a privilege that we often take for granted.

- 5min: what got me into internet archiving

- grew up with unreliable internet
- censored internet
- hostile environment for journalism and content
- discovered wget
- created pocket-archive stream

I just got tired of links I care about 404ing over time

So I built a tool to archive my Pocket stream automatically



**ArchiveBox**  
The open-source self-hosted web archive.

[kstart](#) | [Demo](#) | [Github](#) | [Documentation](#) | [Info & Motivation](#) | [Community](#) |

"Your own personal internet archive" (网站存档 / 爬虫)

IT 🟢 Active last tuesday Stars 5.5k Python >=3.5 Chromium >=59

- v0.0.1 released
- Index links now work without nginx url rewrites, archive can now be hosted on github pages
- added setup.sh script & docstrings & help commands
- made Chromium the default instead of Google Chrome (yay free software)
- added `env-variable` configuration (thanks to <https://github.com/hannah98!>)
- renamed from **Pocket Archive Stream** -> **Bookmark Archiver**
- added **Netscape-format** export support (thanks to <https://github.com/ilvar!>)
- added **Pinboard-format** export support (thanks to <https://github.com/sconeyard!>)
- front-page of HN, oops! apparently I have users to support now 😊?
- added Pocket-format export support

---

- v0.0.0 released: created Pocket Archive Stream 2017/05/05

- history of pocket- archive-stream
- use cases
- personal backup of "your internet"

Then Equifax happened...

**<https://docs.sweeting.me/s/equifax-security-incident>**

To enroll in complimentary identity theft protection and credit file monitoring, click [here](#).

# Cybersecurity Incident & Important Consumer Information

[Consumer Notice](#) [FAQs](#) [Potential Impact](#) [Enroll](#) [TrustedID Premier](#) [Contact Us](#)

## Equifax Releases Details on Cybersecurity Incident, Announces Personnel Changes

September 15, 2017

ATLANTA — As part of the company's ongoing review of the cybersecurity incident announced September 7, 2017, Equifax Inc. (NYSE: EFX) today made personnel changes and released additional information regarding its preliminary findings about the incident.

The company announced that the Chief Information Officer and Chief Security Officer are retiring. Mark Rohwasser has been appointed interim Chief Information Officer. Mr. Rohwasser joined Equifax in 2016 and has led Equifax's International IT operations

© 2017 Equifax. All rights reserved. Equifax is a registered trademark of Equifax Inc. Chief Security Officer: Mr. Andrew

[READ MORE](#)

### Recent Updates

Equifax Releases Details on Cybersecurity Incident, Announces Personnel Changes  
September 15, 2017

A Progress Update for Consumers  
September 14, 2017

A Progress Update for Consumers  
September 13, 2017

A Progress Update for Consumers  
September 11, 2017

Call Center Update  
September 8, 2017

To enroll in complimentary identity theft protection and credit file monitoring, click [here](#).

## Cybersecurity Incident & Important Consumer Information Which is Totally Fake, Why Did Equifax Use A Domain That's So Easily Impersonated By Phishing Sites?

[Consumer Notice](#) [FAQs](#) [Potential Impact](#) [Enroll](#) [TrustedID Premier](#) [Contact Us](#)

Equifax Announces Cybersecurity Incident Involving Consumer Information, Because of Incompetence

**Equifax should have hosted this on equifax.com with a reputable [EV] SSL Certificate.**

**Instead they chose an easily impersonated domain and used a jelly-bean SSL cert that any script kiddie can impersonate in 20min.**

**Their response to this incident leaves millions vulnerable to phishing attacks on copycat sites.**

**This is why you don't put your security incident website on a domain that looks like a scam (with an Amazon SSL cert).**





Only the 2nd mention of wget in NYTimes history.

danso on Sept 21, 2017 [-]

> *Mr. Sweeting explained in his email that a Linux command, "wget," allows anyone to download the contents of a website, "including all images, HTML, CSS, etc."*

According to my research [0], this is the second time in New York Times history that the word "wget" has appeared in the NYT.

The first time was in 2014:

<https://www.nytimes.com/2014/02/09/us/snowden-used-low-cost-...>

<https://docs.sweeting.me/s/equifax-security-incident>

- 5min: equifax story

- equifax breach announced, site launched
- cloned with pocket-archive-stream
- rehosted and forgot about it
- notified of equifax misposts
- goes viral, 2mil hits
- only 2nd mention of wget in NYTimes history

The power of wget is amazing!

```
wget --no-verbose \  
  --adjust-extension \  
  --convert-links \  
  --force-directories \  
  --backup-converted \  
  --span-hosts \  
  --no-parent \  
  -e robots=off \  
  --restrict-file-names=windows \  
  --timeout=60 \  
  --warc-file=archive.warc \  
  --page-requisites \  
  --user-agent="Lalala this is chrome I promise..." \  
  --load-cookies="mycookies.txt" \  
  --compression=auto \  
  --no-check-certificate \  
  --no-hsts \  
  "https://2019.pygotham.org"
```

DEMO TIME

- \* 5 min: Intro to internet archiving tooling
  - \* wget is powerful
  - \* wget has many options and tunables
  - \* heres the ones I chose for ArchiveBox
  - \* demo

## But it's not perfect...

- How do we handle scripts like Javascript?
- How do we handle REST API requests?
- How do we handle dynamic content like video & games?
- Single Page Apps are the worst, plz stop building them
- How do we make sure it works long-term?

### Why is internet archiving hard

- \* Dynamic and interactive content
- \* Private and paywalled content
- \* Content ID and discovery, Base32 is hard
- \* Dealing with the huge amount of data directly vs curating a smaller amount
- \* Archive format longevity tradeoffs (WARC vs html / pdf)

## Enter... headless-browser-based archiving

- History of headless browsers
- Now we can run JS
- [webrecorder.io](#) vs heretrix vs archivebox
- Single Page Apps, games, dynamic content now works
- How do we make sure it works long-term?

## Distributed archiving

- Central authority vs individual archivers?
- Distribute the archiving process or distribute the filesystem?
- Storage concerns, many hard drives are greater than a few
- Spreading content out makes it more resilient
- Can it replace the normal internet?
- Is it something to strive for?

Why is preserving information important? why does humanity create libraries and museums?

\* How has it been done so far?

\* what types of archives end up surviving?

\* What are the benefits of decentralized vs centralized archives?

## Effective archiving

- What happens when the whole internet becomes immutable?
- Is there such a thing as "too much data"?
- Should we be curating what we archive?
- Should people have a right to be forgotten?
- How will it change the culture of the internet?
- How will it impact the long-term durability of the internet?

## How can you get started with archiving today?

- Run ArchiveBox or WebRecorder.io
- Help host / mirror content thats often targeted
  - <https://github.com/pirate/wikipedia-mirror>
  - BitTorrent / I2P
  - DAT / IPFS
- Join the ArchiveTeam
- Donate

<https://github.com/pirate/ArchiveBox/wiki/Web-Archiving-Community>

Archive Wikipedia, TED, project Gutenberg, etc.

<https://github.com/pirate/wikipedia-mirror>

<https://kiwix.org>

<https://other-wiki.zervice.io>



There's a whole community of professionals and volunteers who do this.

<https://github.com/pirate/ArchiveBox/wiki/Web-Archiving-Community>

**Thank You!**  
**Q&A**



[Monadical.com](https://monadical.com) is hiring remote full-stack devs!

**Slides:**

[github.com/pirate/internet-archiving-talk](https://github.com/pirate/internet-archiving-talk)

**Video:**

[youtube.com/watch?v=7eoz\\_EU6-wQ](https://youtube.com/watch?v=7eoz_EU6-wQ)

