

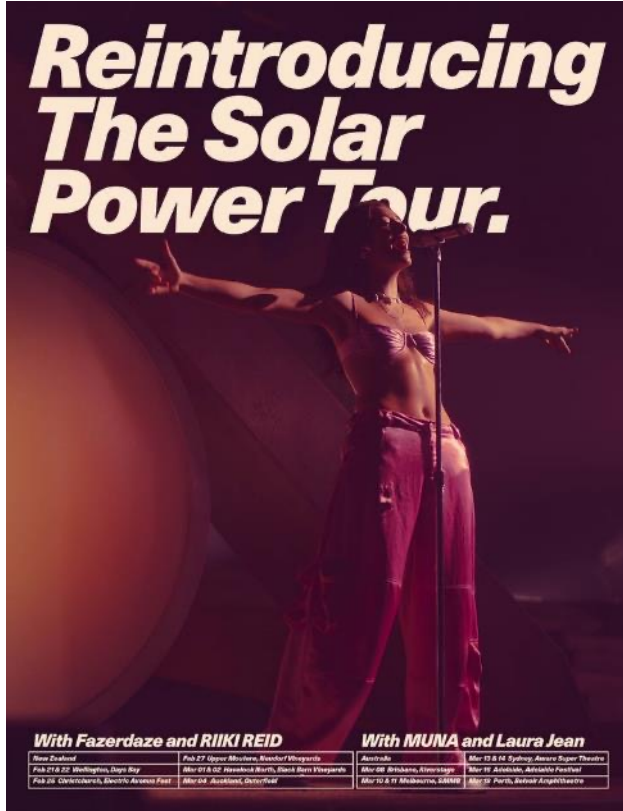


# Browser-based crawling for all: the story so far

Testing Browsertrix at the  
National Library of New  
Zealand







# Collecting on a theme

---

Government

---

Politics

---

Māori

---

Pacific

---

Community Groups

---

Social concerns

---

Music

---

History

---

Arts & culture

---

Sports & recreation

---

Environment

# Significant events - planned

---

Central & Local Body Elections

---

Olympics

---

Commonwealth Games

---

America's Cup

---

Rugby World Cup





If you are sick or have  
symptoms, stay home

Unite  
against  
COVID-19

New Zealand Government

## Significant events – (not so) planned

- Christchurch Earthquakes
- Covid 19
- Christchurch Terror Attack
- Queen Elizabeth II

# Web Curator Tool

- Main web harvesting tool since 2007
- Manages end to end harvesting and archiving process
- CMS for selection decisions, permissions, quality review notes etc.
- limitations in capturing emerging technologies, dynamic content, multimedia and social media, blocked crawls



# We also use these tools

- Archive-it
  - Used for larger websites with multimedia or platforms that we can't capture with WCT
- Conifer (webrecorder)
  - More bespoke harvests of smaller, dynamic sites



# We still couldn't (reliably) capture



- Platforms such as wix, squarespace and shopify
- Embedded social media
- Embedded youtube and other multimedia content
- Infinite scrolling



🔍 Search by name + New Crawl Config

[All](#) [Scheduled](#) [No schedule](#) Show Only Mine  Sort By [Newest](#) ⌵ ⌴

<b>Waiata Māori Music Awards</b> <a href="https://www.waiatamaoriawards.co.nz">https://www.waiatamaoriawards.co.nz</a> 0 crawls <input type="radio"/> No finished crawls <input type="radio"/> No schedule <span style="float: right;">Watch crawl</span>	<b>Mako Road</b> <a href="https://mako.co.nz">https://mako.co.nz</a> 1 crawl 🕒 02/18/23, 1:20 PM <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>The Beths</b> <a href="https://thebeths.com">https://thebeths.com</a> , <a href="https://www.brsa-">https://www.brsa-</a> 1 crawl 🕒 02/10/23, 10:05 AM <input type="radio"/> No schedule <span style="float: right;">Run now</span>
<b>New Zealand String Quartet</b> <a href="https://www.nzsq.org.nz">https://www.nzsq.org.nz</a> 1 crawl 🕒 02/10/23, 9:38 AM <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>Nostalgia Festival</b> <a href="https://nostalgiafestival.co.nz">https://nostalgiafestival.co.nz</a> 1 crawl 🕒 02/09/23, 2:12 PM <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>Pascal Harris</b> <a href="http://www.pascalharris.com">http://www.pascalharris.com</a> 1 crawl 🕒 02/09/23, 10:25 AM <input type="radio"/> No schedule <span style="float: right;">Run now</span>
<b>Sony Music New Zealand</b> <a href="https://www.sonymusic.co.nz">https://www.sonymusic.co.nz</a> 0 crawls <input type="radio"/> No finished crawls <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>Golf New Zealand</b> <a href="https://www.golf.co.nz">https://www.golf.co.nz</a> 0 crawls <input type="radio"/> No finished crawls <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>The Great Kiwi Beer Festival</b> <a href="https://greatkiwibeerfestival.co.nz">https://greatkiwibeerfestival.co.nz</a> 1 crawl 🕒 02/07/23, 2:27 PM <input type="radio"/> No schedule <span style="float: right;">Run now</span>
<b>Otago Cricket Association</b> <a href="https://www.otagocricket.co.nz">https://www.otagocricket.co.nz</a> 1 crawl 🕒 02/07/23, 11:23 AM <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>Play It Strange</b> <a href="https://www.playitstrange.org.nz">https://www.playitstrange.org.nz</a> 1 crawl 🕒 02/07/23, 1:37 PM <input type="radio"/> No schedule <span style="float: right;">Run now</span>	<b>The Slacks</b> <a href="https://theslacks.nz">https://theslacks.nz</a> 1 crawl 🕒 02/03/23, 4:32 PM <input type="radio"/> No schedule <span style="float: right;">Run now</span>

# Browsertrix

- Browsertrix Crawler is a simplified (Chrome) browser-based high-fidelity crawling system, designed to run a complex, customizable browser-based crawl in a single Docker container
- We wanted to test it out on the growing list of sites we haven't been able to harvest

Harvest Issues

- Replay issue - Wombat
- Not picking up secondary news pages
- Dynamic link loading not captured
- Editing crawl templates
- Stuck in stopping
- Infinite scrolling / load more (lazy loading)

+ Add a card

Replay Issues

- Images not loading
- Cannot contact reCAPTCHA. Check your connection and try again. Error in replay.
- iipc.browsertrix.cloud redirected you too many times

+ Add a card

Enhancements

- Ability to delete / remove crawls
- Ability to add notes / annotations to crawls

+ Add a card

Websites to Test

- Sites to test in Browsertrix - Sholto

+ Add a card

Specific Examples

- Age Restricted Pop-ups
- Multiple Seeds - Path Only
- Litho

+ Add a card

Platforms to Test - Add Examples

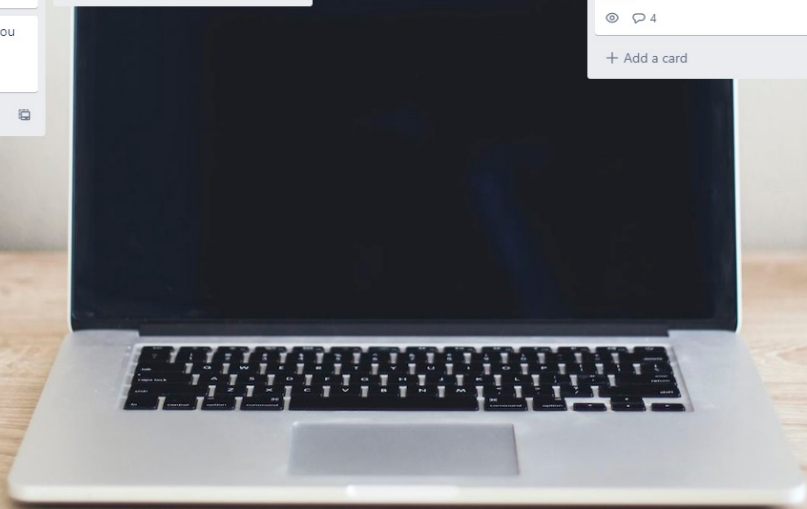
- Squarespace
- Wix
- Shopify
- Sportsground
- Google Docs / Drive
- Externally hosted newsletters etc
- Embedded songkick

+ Add a card

General Questions

- Multiple seeds with different scopeType
- Purge Cache + Full Reload
- Crawl metadata for individual h
- Options for downloading larger collections

+ Add a card



# Wix

- Mixed results in capturing wix sites overall.
- Required some tinkering with crawl config to ensure images were harvested.
- Still issues capturing images from some sites as well as blurred images on playback.
- A number of wix sites we hadn't been able to capture have now been archived



# Squarespace

- Reasonably high success rate.
- Issue capturing images from <https://images.squarespace-cdn.com/>
- Images encoded with intrinsic width values are problematic
- A number of squarespace sites we hadn't been able to capture have now been archived



# Shopify



- Early success with shopify, we were able to capture the Flying Nun website which has been an issue for several years

# Sporty

- Some success in testing sportsground sites, including capturing newsfeeds and dynamic loading content that couldn't capture with other tools.
- Issue with pages that use 'JavaScript void o' to load older news.



# Embedded media

- Embedded youtube videos captured
- Embedded vimeo captured
- Some issues with songkick embed feeds
- Bandcamp embeds unable to be captured, more testing required.



# What we like

Big Flip The Massive

**Crawl Setup** Fields marked with \* are required

Crawl Setup

Browser Settings

Crawl Scheduling

Crawl Information

Confirm Settings

**Crawl Start URL \*** The starting point of your crawl.

**Crawl Scope** Tells the crawler which pages it can visit.

Path Begins with This URL

Will crawl all page URLs that begin with https://bigflipthemassive.weebly.com, e.g. https://bigflipthemassive.weebly.com/path/page.html

**Additional Pages**

**Extra URLs in Scope** Crawl pages outside of Crawl Scope that begin with these URLs.

**Include Any Linked Page ("one hop out")** If checked, the crawler will visit pages one link away outside of Crawl Scope.

**Page Limits**

**Max Pages** Adds a hard limit on the number of pages that will be crawled.

pages

**Exclusions** Specify exclusion rules for what pages should not be visited.

EXCLUSION TYPE	EXCLUSION VALUE

- We can capture a range of websites we previously had issues with
- Easy to set-up crawl configs and run



# What we like

## Crawl of Big Flip The Massive

Scale Stop Cancel ...

Status: **Running** Pages Crawled: 760 / 1,102 Run Duration: 4h 24m 33s Crawler Instances: 1

Overview  
Crawl Queue & Exclusions  
Watch Crawl  
Replay  
Files  
Config

### Queue Exclusions

Exclusions

EXCLUSION TYPE	EXCLUSION VALUE	
Matches Text	bandcamp	🗑
Matches Text	watch	+

• Pending Exclusions: **49 URLs**

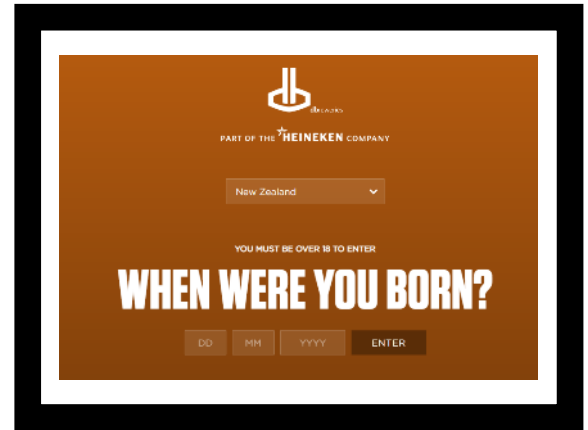
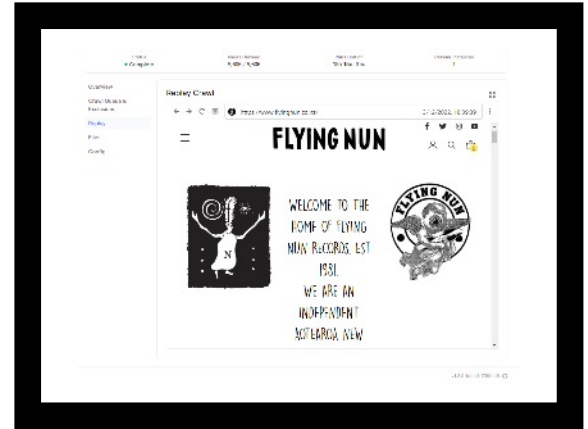
• Crawl Queue: **360 URLs** **101 URLs** < 1 of 12 >

1. [https://www.youtube.com/watch?v=8T52vF4kJs5&embeds\\_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb\\_rel\\_pause](https://www.youtube.com/watch?v=8T52vF4kJs5&embeds_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb_rel_pause)
2. [https://www.youtube.com/watch?v=E\\_Jk2GdF3A5&embeds\\_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb\\_rel\\_pause](https://www.youtube.com/watch?v=E_Jk2GdF3A5&embeds_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb_rel_pause)
3. [https://www.youtube.com/watch?v=gW4z2t6j0s&embeds\\_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb\\_rel\\_pause](https://www.youtube.com/watch?v=gW4z2t6j0s&embeds_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb_rel_pause)
4. [https://www.youtube.com/watch?v=V3gw409FD&embeds\\_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb\\_rel\\_pause](https://www.youtube.com/watch?v=V3gw409FD&embeds_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb_rel_pause)
5. [https://www.youtube.com/watch?v=3TDY07v576m&embeds\\_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb\\_rel\\_pause](https://www.youtube.com/watch?v=3TDY07v576m&embeds_url=https%3A%2F%2Fbigflipthemasive.weebly.com%2F%2Ffeature-emb_rel_pause)
6. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7002287\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7002287_of8.jpg)
7. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/6758151\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/6758151_of8.jpg)
8. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7320805\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7320805_of8.jpg)
9. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/8198207\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/8198207_of8.jpg)
10. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7908889\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/7908889_of8.jpg)
11. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/1246399\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/1246399_of8.jpg)
12. [https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/1246399\\_of8.jpg](https://bigflipthemasive.weebly.com/uploads/1/1/7/2/11426776/1246399_of8.jpg)

- We LOVE being able to watch the crawling live and add exclude filters while the harvest is running

# We also like

- Ability to capture larger websites that were not efficient to manually capture with conifer
- The ability to add browser profiles for sites requiring a login



# Ongoing Challenges

- Issues capturing content that uses lazy loading, this includes infinite scrolling and load more buttons.
- Dynamic link loading not captured which causes playback issues.



# Ongoing Challenges

- Images encoded with intrinsic width values

[https://images.squarespace-cdn.com/content/v1/5e9a373c159e846c7441472e/1587168240228-UHU2TMRIDJO8A78oMM6G/Name\\_preview.png?format=1500w](https://images.squarespace-cdn.com/content/v1/5e9a373c159e846c7441472e/1587168240228-UHU2TMRIDJO8A78oMM6G/Name_preview.png?format=1500w)

The image shows the text "DOLPHIN FRIENDLY" in a stylized, jagged, green font with a black outline. The text is slanted upwards from left to right. The word "DOLPHIN" is on the top line and "FRIENDLY" is on the bottom line.



## Scope

Fields marked with \* are required

Crawl Start URL \*

https://thebeths.com

*i* The starting point of your crawl.

Start URL Scope

Pages on This Domain & Subdomains

*i* Tells the crawler which pages it can visit.

Will crawl all pages on [thebeths.com](https://thebeths.com) and [subdomain.thebeths.com](https://subdomain.thebeths.com).

Include Any Linked Page ("one hop out")

*i* If checked, the crawler will visit pages one link away outside of Crawl Scope.

Exclusions 1

EXCLUSION TYPE	EXCLUSION VALUE	
Matches Text	patreon.com	
+ Add More		

*i* Specify exclusion rules for what pages should not be visited.

Additional URLs 1

List of URLs

https://www.breakfastandtravelupdates.com

*i* The crawler will visit and record each URL listed here. Other links on these pages will not be crawled.

< Cancel

Next >

Save Changes

# What we would like more of

- Easier way to create a crawl config with multiple seeds and different scope types
- Add inclusion rules in a similar way to adding exclusions

# We would like more of

Browsertrix Cloud Sholto Duncan's Archive ▾

[Crawls](#) [Crawl Configs](#) [Browser Profiles](#) [Org Settings](#)

← Back to Crawls

### Crawl of Waiata Māori Music Awards Actions ▾

Status	Pages Crawled	Run Duration	Crawler Instances
● Complete	1,311 / 1,311	4h 2m 34s	1

Overview

**Crawl Queue & Exclusions**

Replay

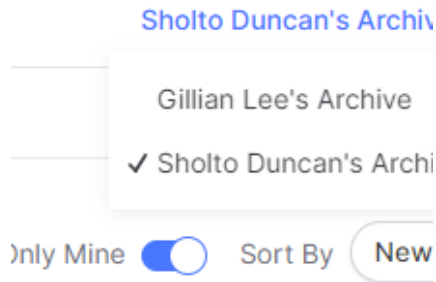
Files

Config

#### Crawl Exclusions

Exclusions	
EXCLUSION TYPE	EXCLUSION VALUE

- Keep crawl queue data post crawl
- Ability to prune crawled content
- Ability to patch crawl or import missing URLs



## We would also like more of

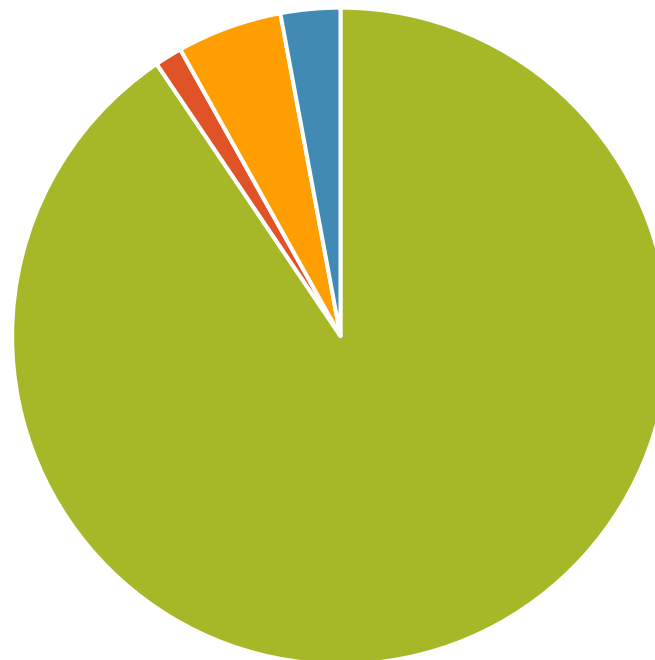
- Compatibility / integration with Web Curator Tool
- Collaboration and sharing of archives within an organisation



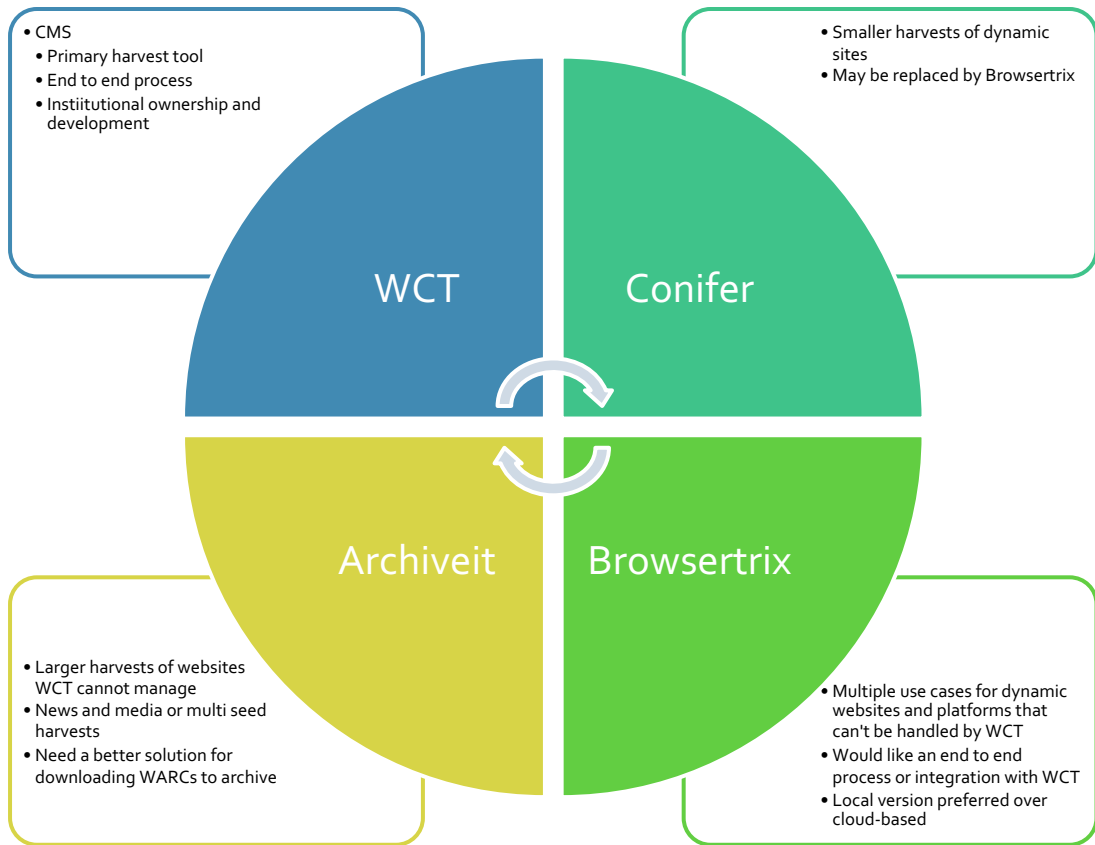


# How browsertrix fits into our Web Archiving program

Content Archived by Web Archiving  
Tools 2022



■ Web Curator Tool ■ Archiveit ■ Conifer ■ Browsertrix



# WEB ARCHIVING TOOLBOX