Browser-based crawling for all: the story so far

Testing Browsertrix at the National Library of New Zealand

Te Kāwanatanga o Aotearoa
New Zealand Government

Te Tari Taiwhenua
Internal Affairs

New Zealand Web Archive

- Legal Deposit legislation and permissions-based
- NLNZ has been collecting websites since 1999
- Active selection since 2007:

    - 38,500+ web harvests archived

    - 9000+ unique website titles

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

Significant events – (not so) planned

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

Web Curator Tool

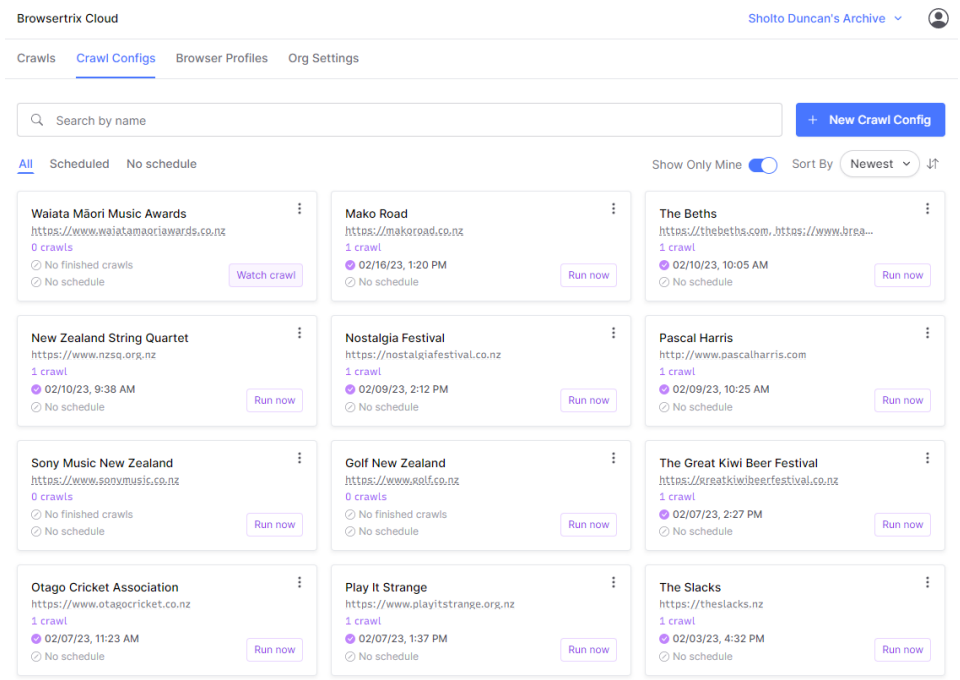
- Main web harvesting tool since 2007
- Manages end to end harvesting and archiving process
- CMS for selection decisions, permissions, quality review notes etc.
- limitations in capturing emerging technologies, dynamic content, multimedia and social media, blocked crawls

We also use these tools

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

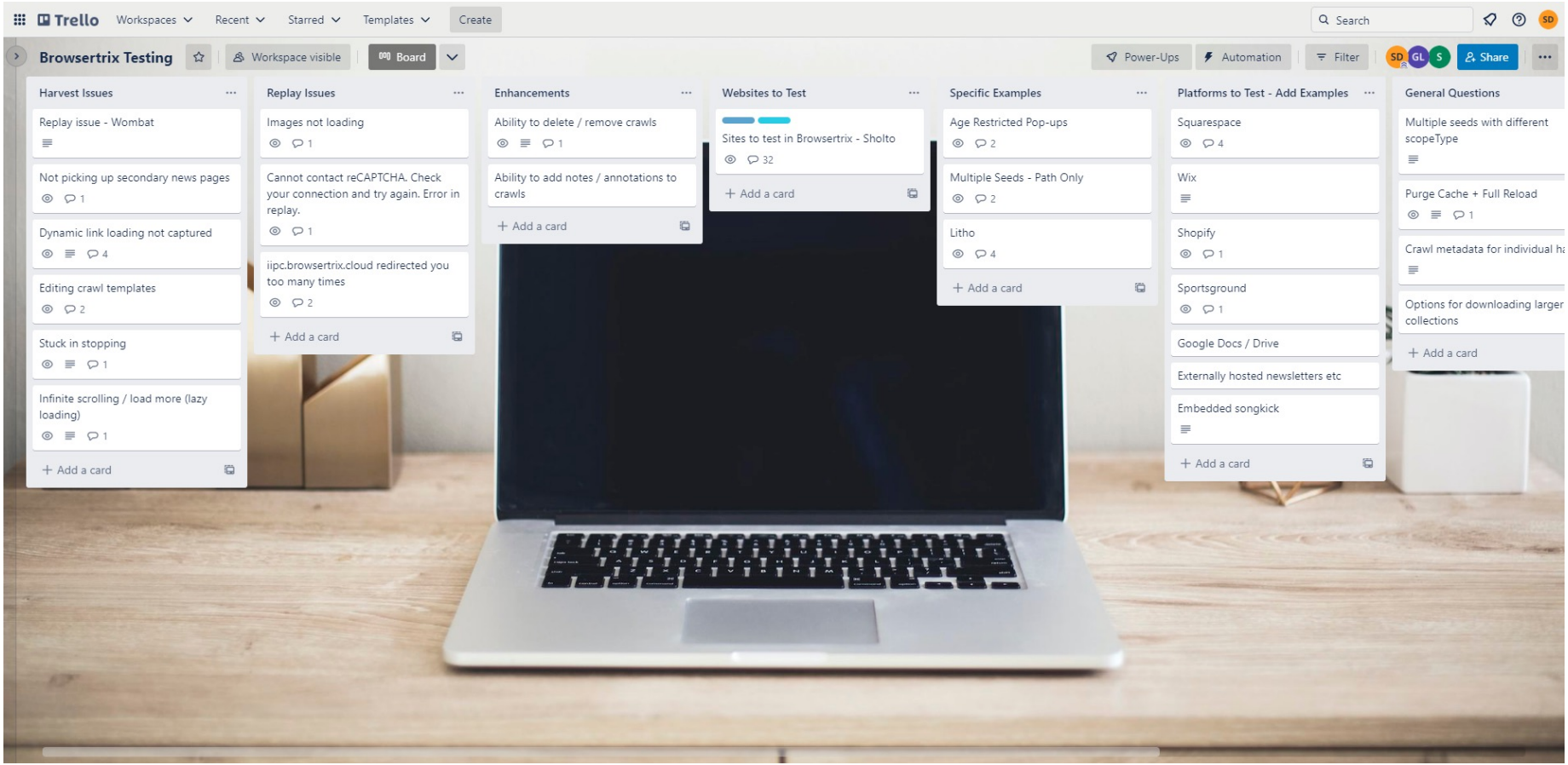
We still couldn't (reliably) capture

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



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


Trello
Workspaces
Recent
Starred
Templates
Create
Search
Browsertrix Testing
Workspace visible
Board
Power-Ups
Automation
Filter
Share
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

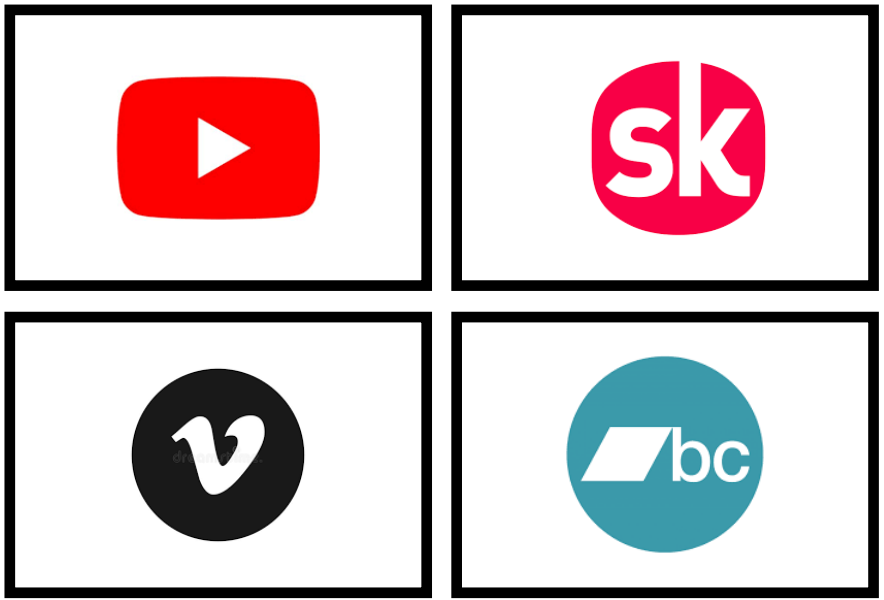
- Early success with shopify, we were able to capture the Flying Nun website which has been an issue for several years

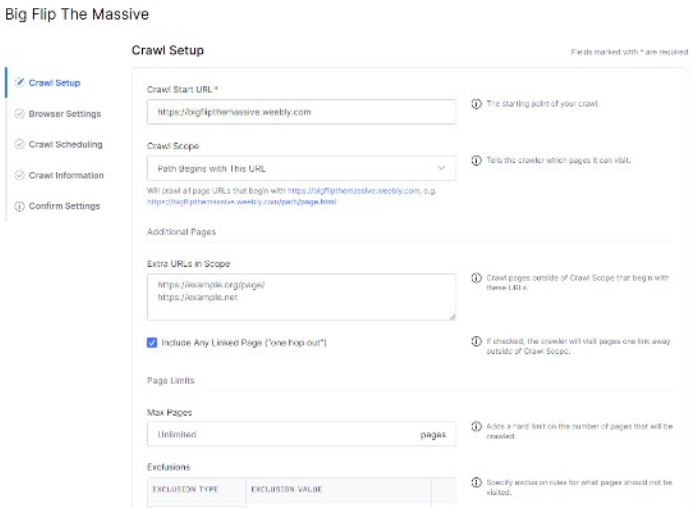
Sporty

- 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

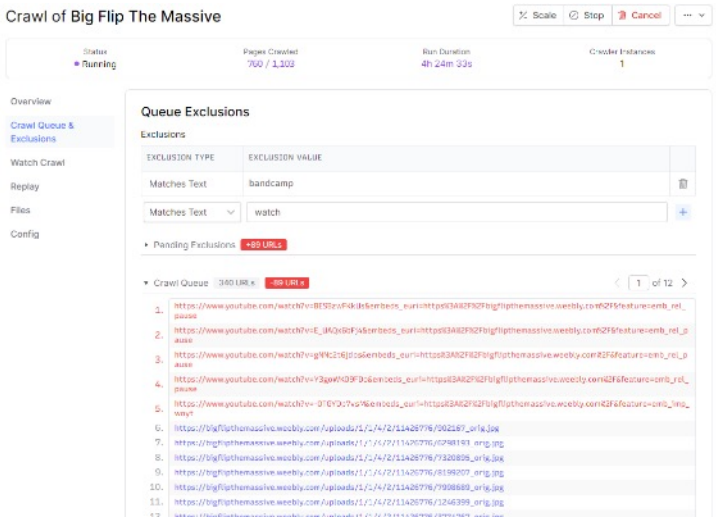
- Embedded youtube videos captured
- Embedded vimeo captured
- Some issues with songkick embed feeds
- Bandcamp embeds unable to be captured, more testing required.





What we like

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



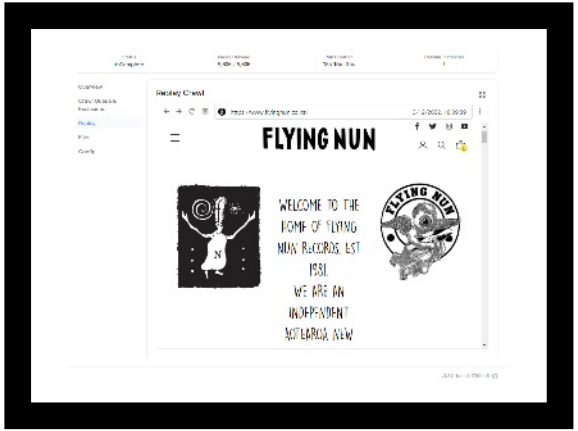
What we like

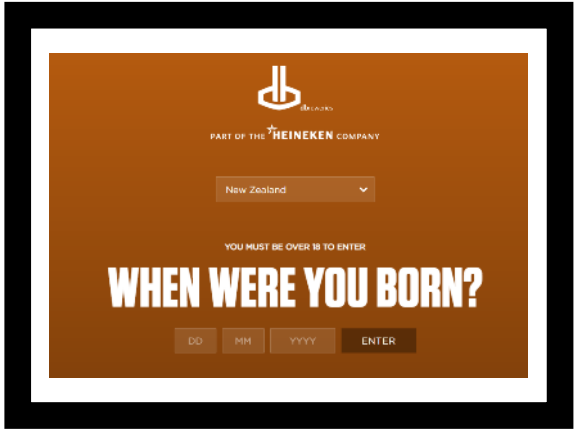
- We LOVE being able to watch the crawling live and add exclude filters while the harvest is running

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.

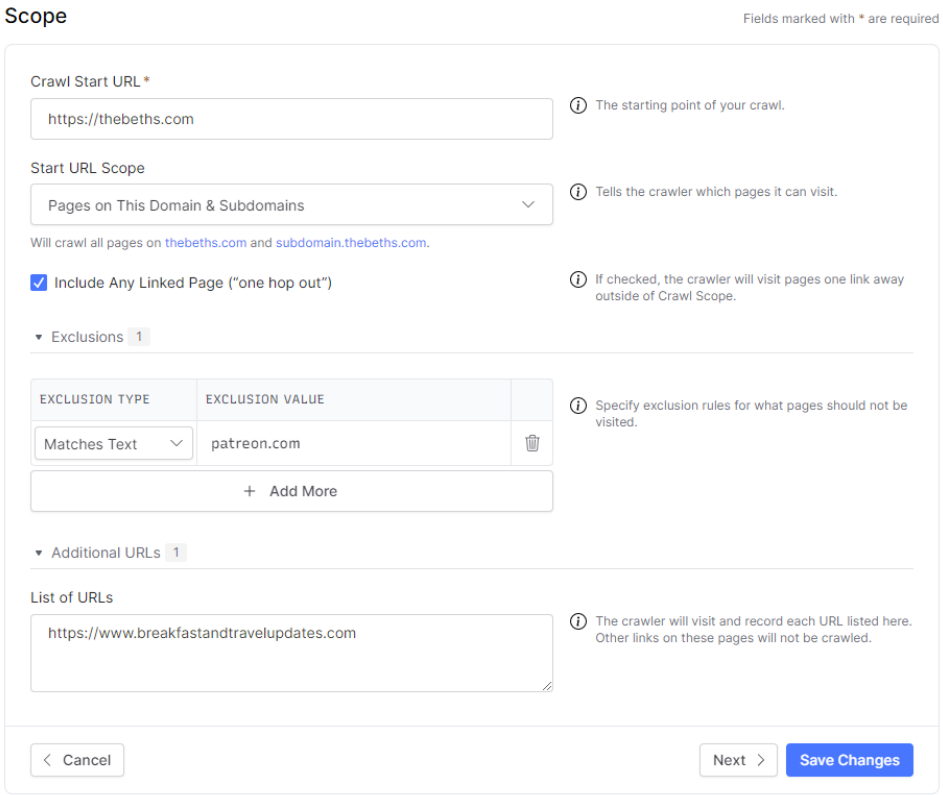
Ongoing Challenges

- Images encoded with intrinsic width values

https://images.squarespace-cdn.com/content/v1/5e9a373c159e846c7441472e/1587168240228-UHU2TMRIDJO8A780MM6G/Name_preview.png?format=1500w

What we would like more of

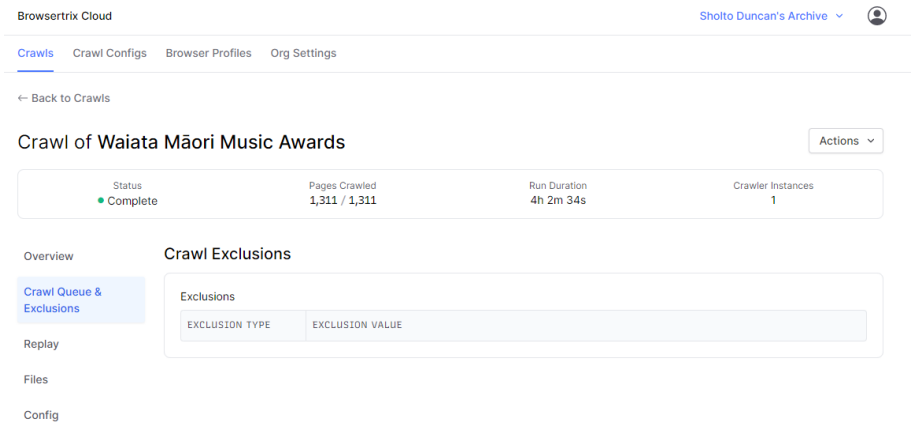
- Ability to delete unwanted crawls
- Add annotations to crawls
- THANK YOU!



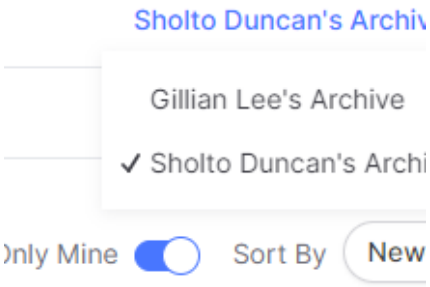
What we would like more of

- 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



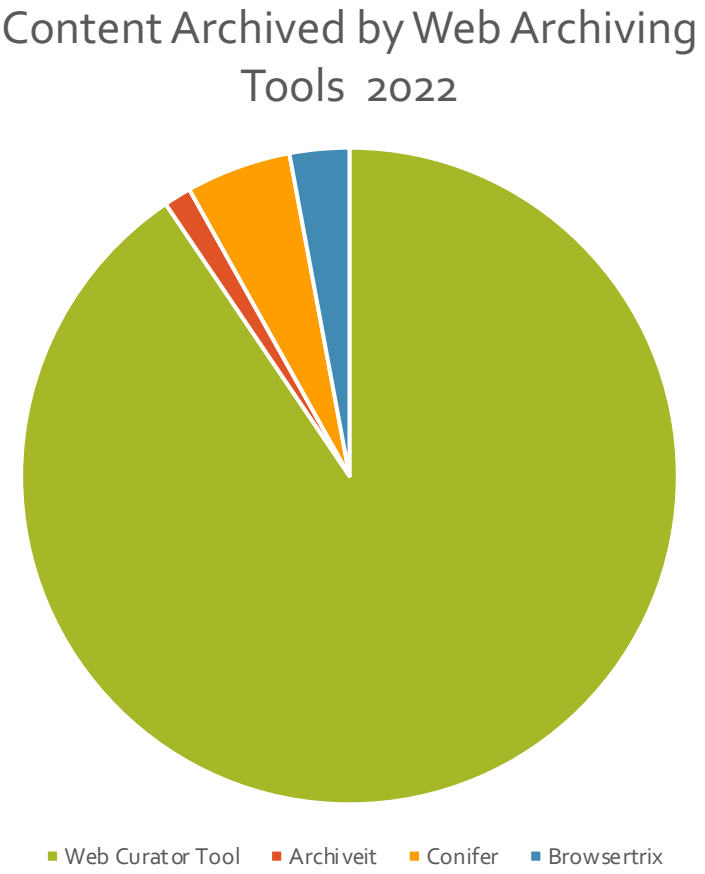
- Keep crawl queue data post crawl
- Ability to prune crawled content
- Ability to patch crawl or import missing URLs



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



WEB ARCHIVING TOOLBOX