# Web Archiving Workflows

This training session was developed in partnership by the International Internet Preservation Consortium (IIPC) and the Digital Preservation Coalition (DPC)

## Workflows We Will Cover

Capture
Preserve
Playback

**Capture:** live web content is downloaded & stored

**Preserve:** downloaded files are checked, converted to a stable file type if necessary, and looked after over time

**Playback:** the archived web content is accessed through a tool that allows users to interact with it like the original

# Capture: Downloading and Storing Live Web Content

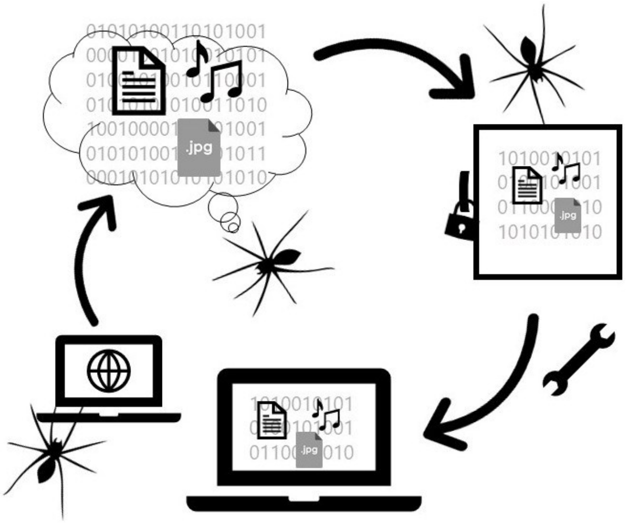
3

## Intro to Crawling

Crawler or “Spider”

Code & files needed to reproduce original website

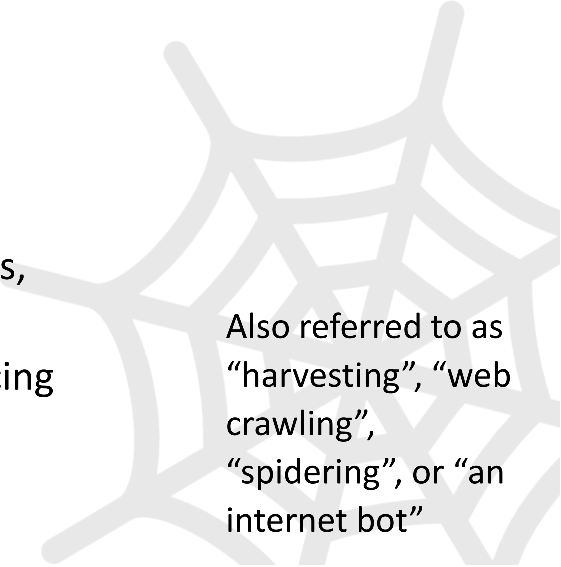
Playback Tool
(like Wayback Machine)



4

# Basic Crawl

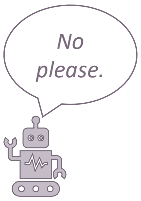
- A tool (‘crawler’) systematically browses the web
- Uses a set of parameters, or defined scope (e.g. from a seed list)
- Downloads code, images, documents, and other files
  - Whatever is essential to reproducing the web content as similarly to original form as possible



Also referred to as “harvesting”, “web crawling”, “spidering”, or “an internet bot”

# Crawling: Parameters

- **Seed URLs** or **URIs**: starting point(s) for web crawler; the crawler follows links out from this initial URL or set of URLs
- **Robots.txt**: a file included in web content that instructs a crawler not to capture that content or to capture only parts of it



- **Crawl Depth** or “**Hops**”: number of links away from the seed URL/URI the crawler will capture

- **Crawl Frequency**: how often the crawler will capture the same URL(s)

# Other Capture Methods

## Dynamic Capture

- Tools built to capture interactive or complex content
  - Ex. videos & other media
  - Ex. social media and other platform-based web
  - Ex. complex JavaScript
- Tools like Webrecorder/Conifer, Internet Archive's Brozzler

## API Harvesting

- Only available for web resources that provide an API
- An API allows authenticated users to extract data directly from the platform through the web
- Works for the modern "platformized" web
- Tools for using APIs: Twarc, Social Feed Manager, others

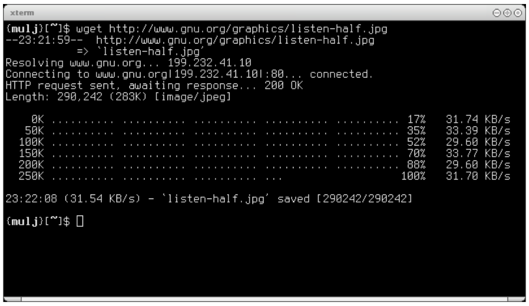
7

# Tools to Support Capture of Web Archives

8

# Tools: GNU wget

- Command line tool that downloads files from the web
- Runs on Unix and Mac OS, but also has a Windows version
- Supports HTTP, HTTPS, and FTP
- Operates continuously in the background
- Usable on slow or unstable networks
- Allows scoping & configuration
- Supports writing to a WARC file
- Free and open source under GNU General Public License



```
{mulj}[~]$ wget http://www.gnu.org/graphics/listen-half.jpg
--23:21:59--  http://www.gnu.org/graphics/listen-half.jpg
           => `listen-half.jpg'
Resolving www.gnu.org... 199.232.41.10
Connecting to www.gnu.org|199.232.41.10|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 290,242 (283K) [image/jpeg]

    0K .......... .......... .......... .......... .......... 17%   31.74 KB/s
   50K .......... .......... .......... .......... .......... 35%   33.39 KB/s
  100K .......... .......... .......... .......... .......... 52%   29.60 KB/s
  150K .......... .......... .......... .......... .......... 70%   33.77 KB/s
  200K .......... .......... .......... .......... .......... 88%   29.60 KB/s
  250K .......... .......... .......... ...                  100%   31.70 KB/s

23:22:08 (31.54 KB/s) - `listen-half.jpg' saved [290242/290242]

{mulj}[~]$ ▯
```

# Tools: Heritrix (+ Umbra)

**Heritrix**

- From the Internet Archive
- Web crawler that downloads websites and embedded media
- Suitable for large collections
- Available for Windows and Unix-like environments
- Supports configurable scoping and deduplication
- Supports writing to a WARC file
- Less effective at triggering and capturing client side script

**Heritrix + Umbra**

- Browser automation tool that runs alongside Heritrix
- Mimics the way a browser would access a page
- Executes client side scripts so previously undetectable URLs can be accessed
- Supports the capture of JavaScript
- Allows for dynamic scrolling

## Tools: Heritrix-Based Curator Tools

- Examples:
  - Archive-It
  - Web Curator Tool
  - NetArchive Suite
- Run Heritrix with user interfaces that make it easier to manage collections
- Many used by IIPC members
- Some curator tools are subscription-based

## Tools: Conifer

- “Browser” + “Crawler” = “Brozzler”
- Captures HTTP traffic as it loads
- Uses a real browser to fetch pages and embedded URLs, and to extract links
- Implements a tool called youtube-dl to improve media capture
- Requires: Python 3.4 or later, RethinkDB deployment, Chromium or Google Chrome version 64 or higher

# Tools: Webrecorder/Conifer

Conifer
by RHIZOME.org
http://conifer.rhizome.org

- User-driven capture rather than automated crawler
- Focus on dynamic web content (embedded video and JavaScript)
- Simple to use interface
- Captures page by page – can be labour intensive!
- Can be structured by Collection and capture session
- Captures can be downloaded as WARC files

**Conifer** = Hosted Service from Rhizome

- Up to 5GB of free storage
- Some use-cases and integrations may require additional support or storage that requires a fee
- Quick Start Guide: https://guide.conifer.rhizome.org/

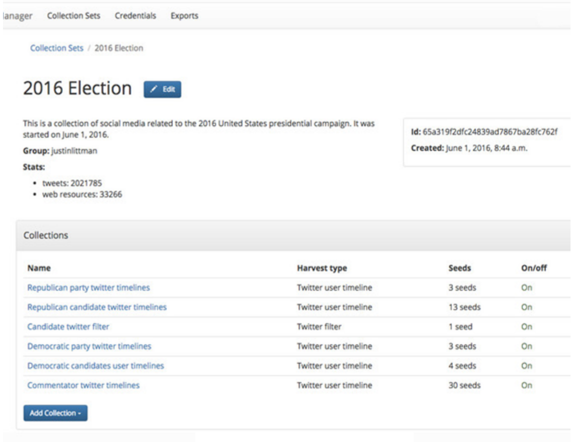
**Webrecorder** = Desktop App

- Same functionality as Conifer but on local desktop
- Slower, but only limited by local storage…

13

# Tools: Social Feed Manager

- Open source tool created by George Washington University Libraries
- Harvests data from Twitter, Flickr, Sina Weibo, and Tumblr
- Captures data through platform APIs
- Captures linked URLs and embedded media
- Supports the curation and management of archived collections



2016 Election

This is a collection of social media related to the 2016 United States presidential campaign. It was started on June 1, 2016.

**Group:** justinlittman

**Stats:**

- tweets: 2021785
- web resources: 33266

**Id:** 65a319f2dfc24839ad7867ba28fc762f

**Created:** June 1, 2016, 8:44 a.m.

Collections

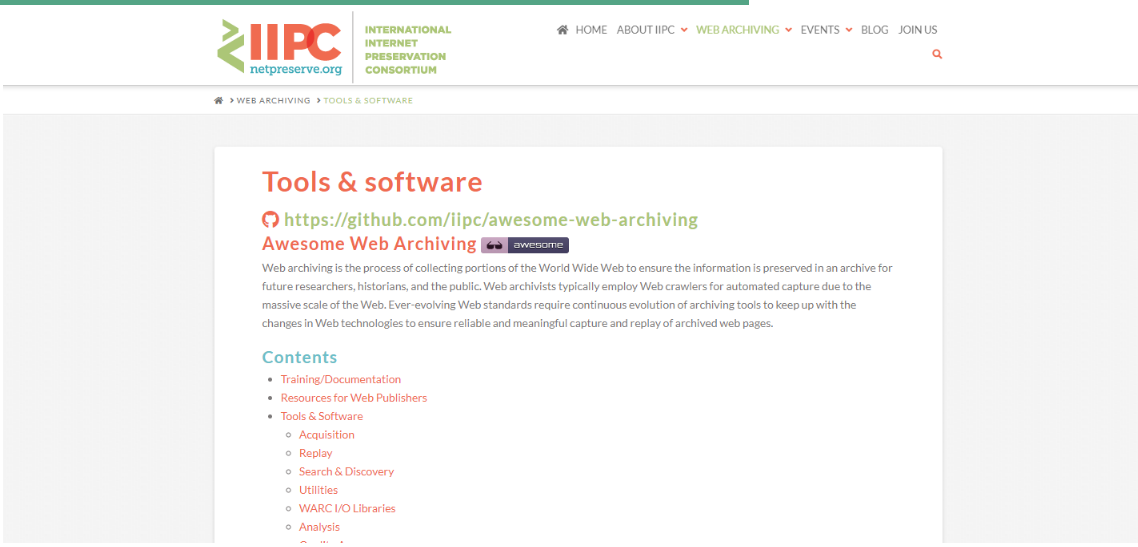
| Name | Harvest type | Seeds | On/off |
|---|---|---|---|
| Republican party twitter timelines | Twitter user timeline | 3 seeds | On |
| Republican candidate twitter timelines | Twitter user timeline | 13 seeds | On |
| Candidate twitter filter | Twitter filter | 1 seed | On |
| Democratic party twitter timelines | Twitter user timeline | 3 seeds | On |
| Democratic candidates user timelines | Twitter user timeline | 4 seeds | On |
| Commentator twitter timelines | Twitter user timeline | 30 seeds | On |

14

# Tools: Social Media Download

- Available for Twitter, Facebook (limited), Google, and others
- Function in Settings
- Only permitted for the account owner
- Good practice for institutions with one or more public-facing social media accounts
- Good practice for personal digital archiving

# Capture: MOAR TOOLS!



## Tools & software

https://github.com/iipc/awesome-web-archiving

Awesome Web Archiving

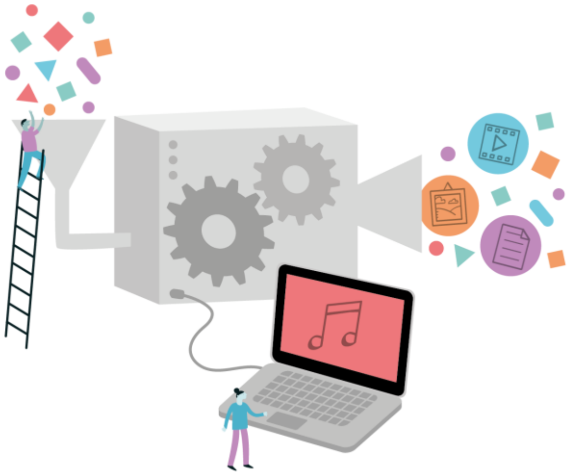
Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. Web archivists typically employ Web crawlers for automated capture due to the massive scale of the Web. Ever-evolving Web standards require continuous evolution of archiving tools to keep up with the changes in Web technologies to ensure reliable and meaningful capture and replay of archived web pages.

Contents

- Training/Documentation
- Resources for Web Publishers
- Tools & Software
  - Acquisition
  - Replay
  - Search & Discovery
  - Utilities
  - WARC I/O Libraries
  - Analysis

https://netpreserve.org/web-archiving/tools-and-software/

# Preserve: What Happens to Captured Content?

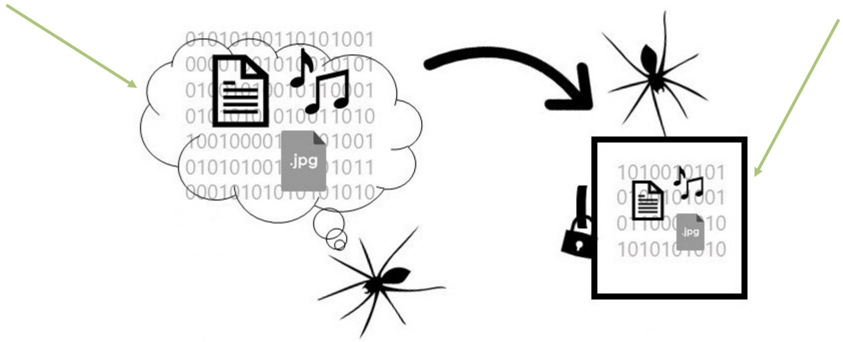


17

## Capture to Preservation

Captured Content
or
"The Crawl"

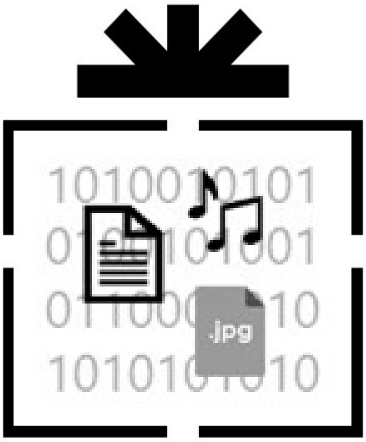
Web Archive
Preserved in Safe
Storage



18

## Introduction to WARC

- WARC (Web ARChive)
- WARC is a wrapper for archived web objects developed by the IIPC
- Tools that write to WARC create files with the extension .warc
- A WARC file can be ingested into a digital preservation system
- WARC was preceded by the ARC (.arc) format



## WARC Standard

- File format standard
- ISO 28500:2017 (formerly ISO 28500:2009)
- Packages together multiple files of different types from a web crawl or capture
- Maintains and describes relationships between web pages or related content
- Accommodates different forms of metadata
- Requires special access tools or viewers

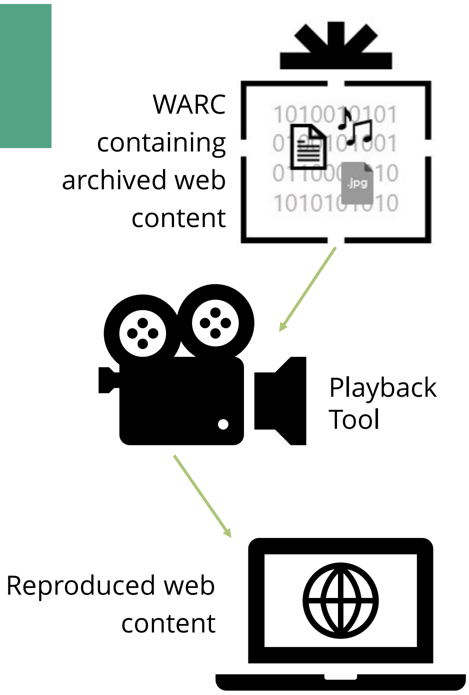
## Preserve: Actions

- Quality assurance
  - Successfully completed capture?
  - Capture content complete?
  - Sensitive data review
- “Patching” any issues
- Generating metadata
- Transfer to archival storage

## Playback: Providing Access to Archived Web Content

# How Do We Provide Access?

- Playback Tools required
- Designed to render archived web content
- Read and display WARC files
- Examples:
  - Wayback Machine
  - ReplayWeb (dynamic, interactive)
  - Third-party service platforms



WARC containing archived web content

Playback Tool

Reproduced web content

23

# Playback: Facilitating Use

24

# Indexing and Search

**Indexing**

- Allows the search and retrieval of archived web content
- Enables search based on metadata fields, including keywords
- Required to enable access & re-use of web content

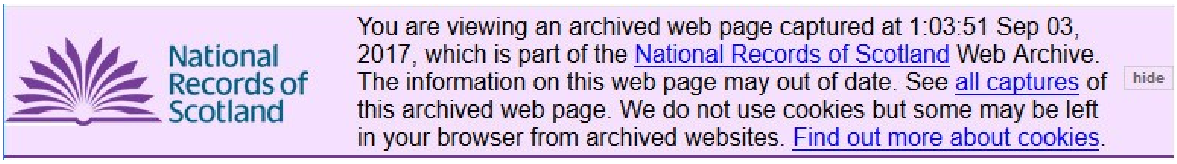
**Full-text Search Index**

- Indexed to allow end users to search broader range of keywords or phrases
- Enhances digital preservation planning
  - UK Web Archive uses full-text search capability to search tags to track the birth and death of specific features like HTML elements

25

# Banner or User Notice

- To designate the viewed content as archived to avoid confusion with the live web



National Records of Scotland

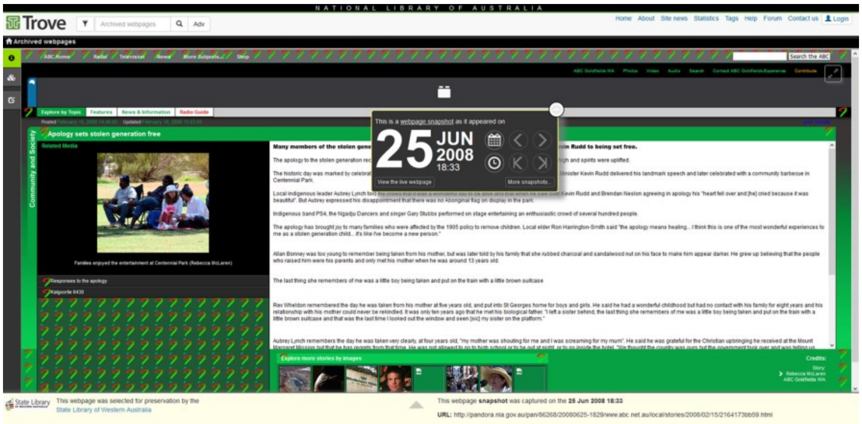
You are viewing an archived web page captured at 1:03:51 Sep 03, 2017, which is part of the National Records of Scotland Web Archive. The information on this web page may out of date. See all captures of this archived web page. We do not use cookies but some may be left in your browser from archived websites. Find out more about cookies.

https://webarchive.nrscotland.gov.uk/#!/

26

# Date and Time of Capture

- To compare with concurrent information, such as major events or other publications



NATIONAL LIBRARY OF AUSTRALIA

Trove | Archived webpages | Adv

Home About Site news Statistics Tags Help Forum Contact us Login

Archived webpages

ABC Home Radio Television News More Subjects... Shop | Search the ABC

ABC Goldfields WA Photos Video Audio Search Contact ABC Goldfields/Esperance Contribute

Explore by Topic Features News & Information Radio Guide

Apology sets stolen generation free

Related Media

Families enjoyed the entertainment at Centennial Park (Rebecca McLaren)

Responses to the apology

Kalgoorlie 6430

This is a webpage snapshot as it appeared on

25 JUN 2008 18:33

View the live webpage | More snapshots...

**Many members of the stolen gene… …vin Rudd to being set free.**

The apology to the stolen generation rec… …gh and spirits were uplifted.

The historic day was marked by celebrat… …Minister Kevin Rudd delivered his landmark speech and later celebrated with a community barbecue in Centennial Park.

Local indigenous leader Aubrey Lynch told the crowd that it was a wonderful day to be alive and that when he saw both Kevin Rudd and Brendan Nelson agreeing in apology his "heart fell over and [he] cried because it was beautiful". But Aubrey expressed his disappointment that there was no Aboriginal flag on display in the park.

Indigenous band PS4, the Ngadju Dancers and singer Gary Stubbs performed on stage entertaining an enthusiastic crowd of several hundred people.

The apology has brought joy to many families who were affected by the 1905 policy to remove children. Local elder Ron Harrington-Smith said "the apology means healing... I think this is one of the most wonderful experiences to me as a stolen generation child... it's like I've become a new person."

Allan Bonney was too young to remember being taken from his mother, but was later told by his family that she rubbed charcoal and sandalwood nut on his face to make him appear darker. He grew up believing that the people who raised him were his parents and only met his mother when he was around 13 years old.

The last thing she remembers of me was a little boy being taken and put on the train with a little brown suitcase

Rex Wheldon remembered the day he was taken from his mother at five years old, and put into St Georges home for boys and girls. He said he had a wonderful childhood but had no contact with his family for eight years and his relationship with his mother could never be rekindled. It was only ten years ago that he met his biological father. "I left a sister behind, the last thing she remembers of me was a little boy being taken and put on the train with a little brown suitcase and that was the last time I looked out the window and seen [sic] my sister on the platform."

Aubrey Lynch remembers the day he was taken very clearly, at four years old, "my mother was shouting for me and I was screaming for my mum". He said he was grateful for the Christian upbringing he received at the Mount Margaret Mission but that he has regrets from that time. He was not allowed to go to high school or to be out at night, or to go inside the hotel. "We thought the country was ours but the government took over and was telling us

Explore more stories by images

Credits:
Story:
Rebecca McLaren
ABC Goldfields WA

State Library — This webpage was selected for preservation by the State Library of Western Australia

This webpage **snapshot** was captured on the **25 Jun 2008 18:33**

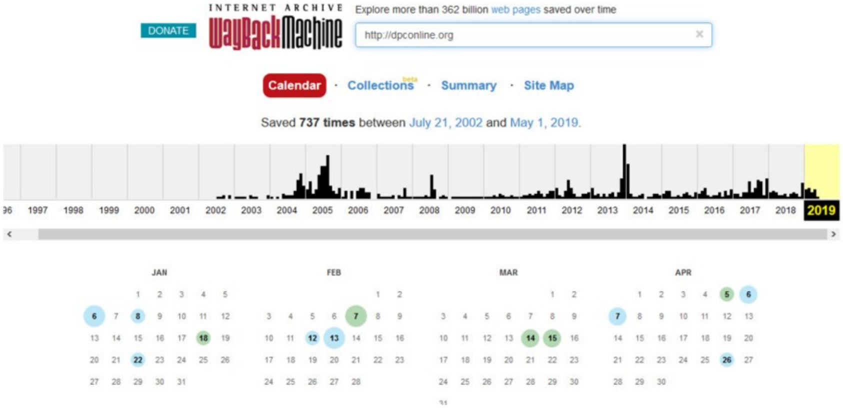
**URL:** http://pandora.nla.gov.au/pan/86268/20080625-1829/www.abc.net.au/local/stories/2008/02/15/2164173bb59.html

https://webarchive.nla.gov.au/awa/20080625083334/http://pandora.nla.gov.au/pan/86268/20080625-1829/www.abc.net.au/local/stories/2008/02/15/2164173bb59.html

27

# Timeline Navigation

- To show the timeline of captures for collections of web content



INTERNET ARCHIVE WayBackMachine | DONATE

Explore more than 362 billion web pages saved over time

http://dpconline.org

Calendar · Collections beta · Summary · Site Map

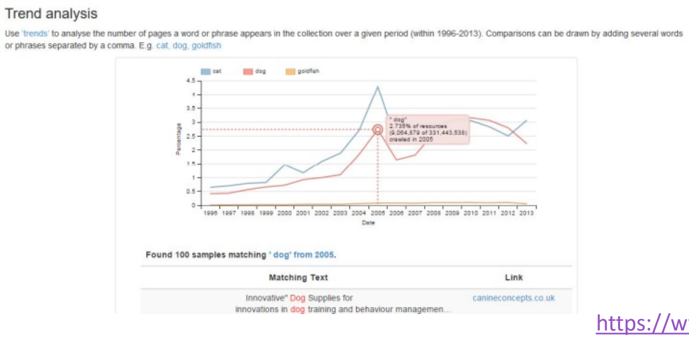
Saved **737 times** between July 21, 2002 and May 1, 2019.

https://web.archive.org/web/*/http://dpconline.org

28

## Web Content as Data

- Users may wish to analyze web content or social media data for trends over time or across the web
- The UK Web Archive’s SHINE historical search engine is a prototype web-based tool for analyzing trends in web content



### Trend analysis

Use 'trends' to analyse the number of pages a word or phrase appears in the collection over a given period (within 1996-2013). Comparisons can be drawn by adding several words or phrases separated by a comma. E.g. cat, dog, goldfish

**Found 100 samples matching ' dog' from 2005.**

| Matching Text | Link |
|---|---|
| Innovative" Dog Supplies for<br>innovations in dog training and behaviour managemen... | canineconcepts.co.uk |

https://www.webarchive.org.uk/shine

## Playback: Tools to Replay Web Archives

# Wayback Machine & OpenWayback

**Wayback Machine**

- Developed by the Internet Archive
- Used to “play back” archived web content contained in a WARC file in an end user’s browser
- Open source software to query and access archived web content

**OpenWayback**

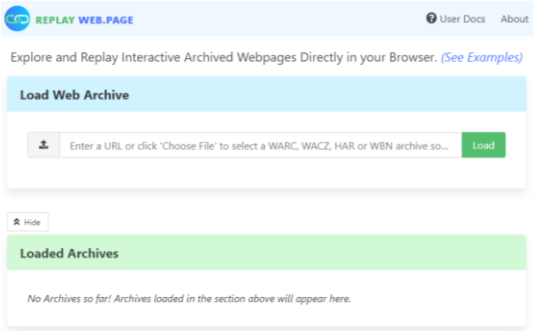
- Shared development project to address common requirements and improve testing

INTERNET ARCHIVE
WaybackMachine

IIPC
netpreserve.org
INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

# ReplayWeb.page

- Developed by the Webrecorder project and complements Conifer and Webrecorder
- Available as a browser-based replay tool or an downloadable app
- Browser-based replay tool can be used offline once it is loaded
- Supports WARC file types (.warc, .warc.gz)



Explore and Replay Interactive Archived Webpages Directly in your Browser.

Load Web Archive

Enter a URL or click 'Choose File' to select a WARC, WACZ, HAR or WBN archive so...

Loaded Archives

*No Archives so far! Archives loaded in the section above will appear here.*

https://replayweb.page/

## Other Tools and Services

- When working with archived web content as data:
  - ArchiveSpark
  - Archives Unleashed Toolkit
- Third-Party Services
  - Archive-It
  - MirrorWeb
  - Hanzo
- For creating Collaborative Collections
  - Momento
  - COBWEB
  - UNT Nomination Tool

33

## Group Discussion

Questions to discuss:

1. Do you currently have web archiving process in place?
2. What tools do you use/have you tested?
3. What would you like to add to your web archiving programme?

34