Introduction to Web Archives

Presented by the web archiving team at the Bodleian Libraries
Objectives

• Gain an understanding of web archives, their importance and potential uses

• Become familiar with UK Web Archive and Bodleian Libraries Web Archive interfaces

• Undertake some exercises using both web archives
# Session plan

<table>
<thead>
<tr>
<th>Introduction to web archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKWA Overview</td>
</tr>
<tr>
<td>Hands-on: UKWA</td>
</tr>
<tr>
<td><strong>Break (15 minutes)</strong></td>
</tr>
<tr>
<td>BLWA Overview</td>
</tr>
<tr>
<td>Hands-on: BLWA</td>
</tr>
<tr>
<td>Further Resources</td>
</tr>
<tr>
<td>Hands-on: Further Resources</td>
</tr>
<tr>
<td>Questions and feedback</td>
</tr>
</tbody>
</table>
Introduction to Web Archives

• What is 'Web Archiving'?

• The process of capturing and preserving websites.

• Websites are crawled by a web crawler.

• The crawler saves crawls in the WARC (Web ARChive) format.

• The same websites are usually web archived periodically (e.g. monthly, yearly)
Introduction to Web Archives

• What is a 'Web Archive'?  
  • A collection of archived websites usually stored in a WARC (Web ARChive) format.  
  • Can be accessed through an access platform  
  • Websites can be arranged thematically or as a part of a bulk capture.  
  • When the information on a web archive is downloaded they become datasets for researchers.
The value of web archives / Why should we web archive?

- Legal / Evidential value:
  - To ensure government, organisational and public accountability
  - Proof that a website existed, websites constantly change and disappear

- Cultural value:
  - To collect valuable websites that form part of our cultural heritage
  - Provide research tools to current / future generations
Page not found (404)

Request Method: GET
Request URL: http://www.jammaplus.co.uk/forum/default.asp/

Using the URLconf defined in isomonit.urls, Django tried these URL patterns, in this order:

1. ^accounts/login/$
2. ^accounts/logout/$
3. ^accounts/profile/$
4. ^admin/$
5. ^devices$
6. ^device/(\?P<device_id>[d-]+)/setup$
7. ^device/(\?P<device_id>[d-]+)/setup/name$
8. ^device/(\?P<device_id>[d-]+)/setup/input/(\?P<input_id>[d-]+)$
9. ^device/(\?P<device_id>[d-]+)/monitor$
10. ^device/(\?P<device_id>[d-]+)/monitor/core$
11. ^locations$
12. ^location/add$
13. ^location/(\?P<location_id>[d-]+)/mod$
14. ^location/(\?P<location_id>[d-]+)/setup$
15. ^location/(\?P<location_id>[d-]+)/setup/input/(\?P<input_id>[d-]+)$
16. ^contact-groups$
17. ^contact-group/add$
18. ^contact-group/(\?P<group_id>[d-]+)/setup$
19. ^contact-group/(\?P<group_id>[d-]+)/setup/name$
20. ^contact-group/(\?P<group_id>[d-]+)/dc$
21. ^contact-group/(\?P<group_id>[d-]+)/add_contact$
22. ^contact-group/(\?P<group_id>[d-]+)/del_contact/(\?P<contact_id>[d-]+)$
23. ^contact-group/(\?P<group_id>[d-]+)/edit_contact/(\?P<contact_id>[d-]+)$
24. ^tec5$
25. ^tec5cmdacks$
26. ^doc/description$
27. ^about$
28. ^$ 

The current URL, forum/default.asp/, didn't match any of these.

You're seeing this error because you have DEBUG = True in your Django settings file. Change that to False, and Django will display a standard 404 page.

Source: UKWA, Online Enthusiast collection
UK Web Archive: background and access

Legal Deposit Libraries Act (2003) NPLD, and further enabling regulations to include websites(2013)

UK top level domain without seeking permission – only accessible on dedicated legal deposit library terminals. Unless the site is Open Access.

Websites in the UK web archive are 'intended to be reflective of contemporary UK life or for their relevance to a particular collection. Judgement is not made on the validity or quality of website content'. UK Web Archive (2018)

2017 data = 500TB

Is the UKWA being safeguarded for future use? - YES!
UK Web Archive: collections and our role
Demo: how to search and navigate the UKWA

Point browser to https://www.webarchive.org.uk
Explore over 100 collections of websites brought together by librarians, curators and other experts in response to a wide range of events and diverse topics and themes.

19th Century English Literature
Resources relating to 19th century English literature and...

Aging
This collection looks at aging issues from a number of co...

Black and Asian Britain
Collection focussing on Black and Asian communities and c...

Blogs
The UK Blogosphere (connected community of Web logs) has ...
Black Dog Tribe (@FollowBDT) on Twitter

Online community for anyone affected by mental illness, r...

Blurt Foundation (@Blurtitout) on Facebook

A facebook page and an online place to talk openly about ...

Blurt: INCREASING AWARENESS AND UNDERSTANDING OF DEPRESSION

BoBfm: Welwyn Hatfield MP calls for more pro-active approach to suicide prevention
#HOMETOVOVOTE

ARE YOU IRISH & LIVING ABROAD?

FIND OUT HOW YOU CAN HELP REPEAL THE 8TH.

I LIVE IN IRELAND  I LIVE ABROAD
# Search Results

21 captures of [https://hometovote.com/](https://hometovote.com/)

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018</td>
<td>March</td>
<td>1 March</td>
<td>06:49:56</td>
</tr>
<tr>
<td></td>
<td>April</td>
<td>1 April</td>
<td>13:01:08</td>
</tr>
<tr>
<td></td>
<td>May</td>
<td>22 May</td>
<td>15:09:07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22 May</td>
<td>16:55:24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26 May</td>
<td>00:37:14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29 May</td>
<td>14:07:51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29 May</td>
<td>15:45:09</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>4 June</td>
<td>06:49:30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 June</td>
<td>10:31:59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 June</td>
<td>09:26:19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 June</td>
<td>10:02:47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18 June</td>
<td>09:41:33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18 June</td>
<td>10:30:05</td>
</tr>
</tbody>
</table>
Search results: 44,863 results for "blog"

Sort by: Most relevant

One World Trust

Viewable only on Library premises

http://www.oneworldtrust.org/

Save a UK website

Fields marked with * are mandatory

Personal details you provide on this form are protected by UK Data Protection Law. Please view our Privacy Statement.

Name
Please enter your name

* Email address [*]
Please enter your email address

* Title of website
Please enter the title of website

* URL of the website
Please enter URL of website

Any other information
Use this space to tell us for example, if you own this website, or if it is about to go offline

[*] we need your email address if you require a response from us. It would also be helpful if you are the owner of the nominated website and would like us to make archival copies of the website available through the UK Web Archive.

I'm not a robot

Submit
UK Web Archive: Worksheet

- Work your way through the UK Web Archive worksheet.
- If you need any help please raise your hand.
Bodleian Libraries' Web Archive

- Started in 2011
- Hosted by Archive-It
- Contains over 11.5 TB of data and 167,416,167 docs
- Focuses on capturing sites related to Oxford and current Bodleian Special Collections.
BLWA Collecting areas:

- Arts and Humanities
- UO Museums, Libraries and Archives
- Social Sciences
- International
- University of Oxford
- Oxford Colleges
- Science, Medicine and Technology
- UO Student societies
Demo: how to search and navigate the BLWA

Point browser to https://archive-it.org/home/bodleian
Bodleian Libraries
Archive-it Partner Since: Dec, 2010
Organization Type: Colleges & Universities
Organization URL: http://www.bodleian.ox.ac.uk/
Description: The Bodleian Libraries is the integrated library service of the University of Oxford. Established in 2000, it comprises nearly 40 libraries. Among these are major research libraries - including the Bodleian Library, which has been a library of legal deposit for almost 400 years - as well as libraries attached to faculties, departments and other institutions of the University. The combined collections of the Bodleian Libraries number more than 11 million printed items, and include extensive special collections and manuscripts.

Narrow Your Results

Sites and collections from this organization are listed below. Narrow your results at left, or enter a search query below to find a collection, site, specific URL or to search the text of archived webpages.

Enter search terms here

Collections | Sites | Search Page Text

Page 1 of 1 (7 Total Results)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

Arts and humanities
Archived since: Mar, 2011
Description: Web sites relating to arts and humanities.
Subject: Arts & Humanities

International
Archived since: Mar, 2011
Description: Web sites from around the world.
Subject: Arts & Humanities, Politics & Elections, Society & Culture
Title: The Agrippa Files

URL: http://agrippa.english.ucsb.edu/

Description: a scholarly site that presents archival documents, commentaries and study tools on the original art book, 'Agrippa (a book of the dead)'

Captured 14 times between Jan 31, 2012 and Jan 4, 2019

Subject: an online archive of Agrippa (a book of the dead)

Creator: Transcription Project by the English department at the University of California, Santa Barbara

Publisher: UC Santa Barbara Transcriptions Project and University of California, Santa Barbara

Language: English
Found 14 Captures between Jan 31, 2012 - Jan 4, 2019

<table>
<thead>
<tr>
<th>Year</th>
<th>Pages</th>
<th>Date</th>
<th>Page Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>5 pages</td>
<td>Jan 31, 2012</td>
<td>*</td>
</tr>
<tr>
<td>2013</td>
<td>2 pages</td>
<td>Apr 19, 2013</td>
<td>*</td>
</tr>
<tr>
<td>2014</td>
<td>2 pages</td>
<td>Apr 19, 2014</td>
<td>*</td>
</tr>
<tr>
<td>2015</td>
<td>1 page</td>
<td>Apr 21, 2015</td>
<td>*</td>
</tr>
<tr>
<td>2016</td>
<td>1 page</td>
<td>Jan 4, 2016</td>
<td>*</td>
</tr>
<tr>
<td>2017</td>
<td>1 page</td>
<td>Jan 4, 2017</td>
<td>*</td>
</tr>
<tr>
<td>2018</td>
<td>1 page</td>
<td>Jan 4, 2018</td>
<td>*</td>
</tr>
<tr>
<td>2019</td>
<td>1 page</td>
<td>Jan 4, 2019</td>
<td>*</td>
</tr>
</tbody>
</table>
“I hesitated before untying the bow that bound this book together...”

Agrippa (a book of the dead) appeared in 1992 as a collaboration between artist Dennis Ashbaugh, author William Gibson, and publisher Kevin Begos, Jr. The Agrippa Files is a scholarly site that presents selected pages from the original art book; a unique archive of materials dating from the book's creation and early reception; an emulation of Gibson's included poem in its original born-and-die-digital form (it ran from a diskette once before an
Bodleian Libraries' Web Archive: Worksheet

• Work your way through the BLWA worksheet.
• If you need any help please raise your hand.
Potential issues on BLWA

Content not captured

Dynamic content

Videos: Youtube / Vimeo

Not in Archive

The page you requested has not been archived in Archive-It.

This could be for a number of reasons.

Most likely the page you are requesting was outside of the crawler's scope. Try another request or click to see other pages from [https](https://).
Extra tools and resources

Time Travel – check if other web archives have captured a site - http://timetravel.mementoweb.org

Tamper Data Firefox or Chrome browser extension – ability to find missing elements of a web page
Demo: Voyant Tools

Point browser to
https://voyant-tools.org
Try inputting 'https://www.richarddawkins.net/' and clicking 'reveal'
hide You are viewing an archived web page, collected at the request of Bodleian Libraries using Archive-It. This page was captured on 11:10:33 Aug 10, 2018, and is part of the Science, medicine and technology collection. The information on this web page may be out of date. See All versions of this archived page. Enable QA View Missing URLs

CLOSE

About What We Do Richard Dawkins Merger with CFI FAQ Board Staff Press Website FAQ News Community Paragraph of the Week The Hotline Project The Clergy Project Discussions Archive Convert's Corner Mailbox Tour Videos
SHINE interface: Trends and searching

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

[Further information about the dataset, including how to cite: http://data.webarchive.org.uk/opendata/ukwa.ds.2/#issues]
<table>
<thead>
<tr>
<th>Matching Text</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>collections Overview British Business</td>
<td><a href="http://www.nls.uk">www.nls.uk</a></td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>Domarker tool Low volume applications WinCLime in Ireland Condensing vacuum steam For classic electrophoresis Next gener contamination Electrophoresis video New level of performance Hg New solutions program Calibrate micropipettes New objectives Accurate volume PCR reactions MinioRNA components Lab for ancient DNA tomography X-ray microanalysis system Family for sample prep smarter August - 2007</td>
<td>labnews.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
<tr>
<td>distracting. Quite a lot of care has been taken in this firm's newsreels and their preservation, and it shows. The sound is a different matter: orig… provided by Screensound Australia. It runs ten minutes and is…</td>
<td>dvdtimes.co.uk</td>
</tr>
</tbody>
</table>
Extra Tools: Worksheet

- Work your way through the Extra tools worksheet.
- If you need any help please raise your hand.
Other Web Archives

Government archive
http://www.nationalarchives.gov.uk/webarchive

Internet archive
https://archive.org/

Parliamentary archive
http://webarchive.parliament.uk

Other institutional archives
https://archive-it.org/explore?show=Organizations
Further information on web archives:

- Web Archive LibGuide
  - More information about BLWA and UKWA
  - Other examples of web archives
  - Web archive use cases
  - Other web archiving tools

URL: https://libguides.bodleian.ox.ac.uk/web-archives
Thank you for your attention

Questions?