

UK Web Archive

DPC Webinar: Approaches and Tools – scalable and collaborative web archiving

Nicola Bingham

Lead Curator, Web Archiving
[@NicolaJBingham](#)

Boston Spa
30th January 2020

UK Web Archive Approaches and Tools – scalable and collaborative web archiving

- 1) Background
- 2) Legislative framework
- 3) Domain crawling
- 4) Selective crawling
- 5) Content Strategies
- 6) Partnerships and Collaborations



British Library, Boston Spa

UK Web Archive Background

Latest Instances | Translate to Welsh

UK WEB ARCHIVE
preserving uk websites

Archived August 2005 Archived November 2005 Archived May 2006 Archived June 2007 Archived March 2009
Archived October 2004 Archived March 2005 Archived November 2006 Archived November 2008 Archived May 2009

You are here: Home

Provided by: **LIBRARY HSIILIB**

Welcome to the UK Web Archive
Thousands of UK websites have been collected since 2004 and the Archive is growing fast.

Here you can see how sites have changed over time, locate information no longer available on the live Web and observe the unfolding history of a spectrum of UK activities represented online. Sites that no longer exist elsewhere are found here and those yet to be archived can be saved for the future by nominating them.

The Archive contains sites that reflect the rich diversity of lives and interests throughout the UK. Search is by Title of Website, Full Text or URL, or browse by Subject, Special Collection or Alphabetical List.

Cookies on the UK Web Archive website
We use cookies to ensure that we give you the best experience on our website. If you continue without changing your settings, we'll assume that you are happy to receive all cookies on the UK Web Archive website. [Click here for more information.](#)

Quick links
What is the UK Web Archive?
Who is the UK Web Archive for?
How do I search the archive?
How can I nominate a website?
Video: An Introduction to the UK Web Archive (7 minutes).

Browse by Subject
Arts & Humanities
Business, Economy & Industry
Education & Research
Government, Law & Politics
Medicine & Health
Science & Technology
Society & Culture
Browse by Subject

Quick search
Please enter text
Full text (across all the archived websites)
Title (for a specific archived website)
search
Advanced search

Explore the Special Collections
Special Collections are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field. They can be events-based (e.g. The Olympic & Paralympic Games 2012), topical (e.g. The Credit Crunch Collection) or subject-oriented (e.g. The British Countryside Collections).

Blogs Credit Crunch Live Art London Terrorist Att... Olympic & Paralympic...
Personal Experiences... Quakers Queen's Diamond Jubl... UK General Election ... UK General Election ...

Browse Special Collections

Visualisations
N-gram Search

Notice and takedown | Terms and conditions | Privacy statement

- 2004 began selectively archiving websites
- Legal Deposit Libraries Act 2003
- Permission sought before archiving
- 30% success rate
- Resource intensive
- 2004 - 2013 = 15,000 websites

UK Legal Deposit Libraries



- British Library St Pancras
- British Library Boston Spa
- National Library of Scotland
- National Library of Wales
- Trinity College Dublin
- Bodleian Libraries, Oxford University
- Cambridge University Library



*Server racks,
Boston Spa*

UK Web Archiving Team



Non- Print Legal Deposit Regulations

Definition of a “UK work”:

- a) It is made available to the public from a website with a domain name which relates to the UK; or
- b) Is made available to the public by a person and any of that person’s activities relating to the creation or the publication of the work take place within the United Kingdom.

[The Legal Deposit Libraries (non-print works) regulations, 2013]

Automatically scope in:

- UK TLDs, e.g. .uk, .scot, .cymru, .london etc
- Geo-IP database look up for UK servers



Out of scope / Exclusions

- Film and recorded sound where the audio-visual content predominates, e.g. YouTube, BBC iPlayer.
- Private intranets and emails.
- Personal data in social networking sites or that are only available to restricted groups.

Domain Crawling

Annual Domain Crawl average stats:

- 5-10 million hosts (websites)
- Over 2 billion items
- 70 - 100 TB of compressed data

NPLD totals

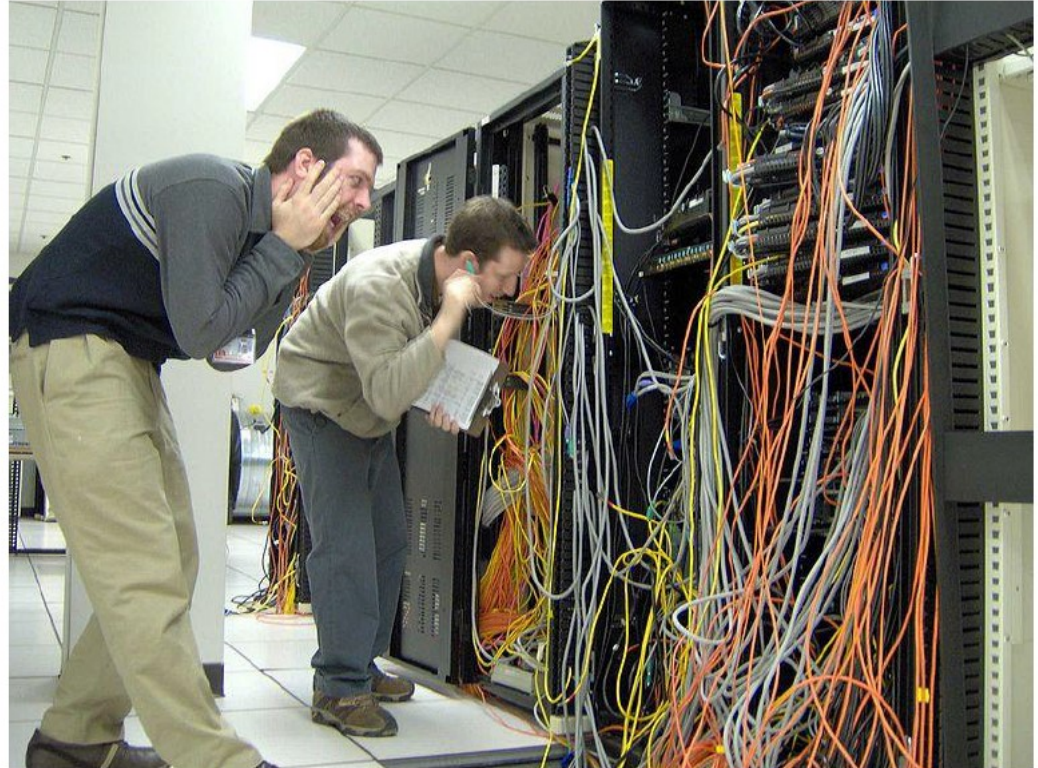
Domain	477.64 TB
Frequent	255.24 TB
Total	732.88 TB
Webrecorder	1.63 TB

" [The UK Web Archive contains] *shocking amounts of information*"

Milligan, I. (2015) Web Archive Legal Deposit: A Double-Edged Sword. Digital History, Web Archives, and Contemporary History. <https://ianmilligan.ca/2015/07/14/web-archive-legal-deposit-a-double-edged-sword/> 14th July 2015 [Date accessed: 09/07/2018]

There are (many) technical limitations

- Database driven sites
- Programming Scripts
- Plug-Ins
- Proprietary file formats
- Blockers
 - Robots.txt
 - Access denied



Annotation, Curation Tool

Targets > List

▼ Filters ▲

Apply

Clear

Export

Curator

None

Organisation

None

Crawl frequency

None

Depth

None

Licence

None

Flag

None

Page size

10

Subject

Collection

- ☐ Arts & Humanities
 - ☐ Business, Economy & Industry
 - ☐ Education & Research
 - ☐ Government, Law & Politics
 - ☐ Medicine & Health
 - ☐ Science & Technology
 - ☐ Society & Culture
-
- ☐ 100 Best Sites
 - ☐ 19th Century English Literature
 - ☐ 42 Days
 - ☐ Aging
 - ☐ Arnhem 75
 - ☐ Black and Asian Britain
 - ☐ Blogs
 - ☐ Brexit
 - ☐ British Countryside
 - ☐ Britishness
 - ☐ British Overseas Territories
 - ☐ British Stand-up Comedy Archive
 - ☐ Cambridge Network
 - ☐ Cambridge Science
 - ☐ Canada UK
 - ☐ Capeli ac Eglwysi Cymreig/ Welsh Churches
 - ☐ Caribbean Communities in the UK
 - ☐ Celtic Studies
 - ☐ Children's Websites
 - ☐ Climate Change Debates
 - ☐ Conservative Party Website deletions - Pre
 - ☐ Cornwall
 - ☐ Credit Crunch
 - ☐ Crimean War
 - ☐ Darwin 200

Topics and Themes (Special Collections)



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 ...



Brexit

Brexit is an abbreviation for "British exit," referring t...



British Countryside

Collection of Internet sites selected by staff at the Bri...



British Overseas Territories

This is a collection of archived websites related to the ...



British Stand-up Comedy Archive

The British Stand-Up Comedy Archive was established at th...


Who Curates the Web Archive?




Trinity College Dubli
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin




Collaboration IIPC



[HOME](#)
[EXPLORE](#)
[LEARN MORE](#)
[CONTACT US](#)

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive


Explore >> International Internet Preservation Consortium



INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM
iipc.org

International Internet Preservation Consortium

Archive-IT Partner Since: Dec, 2013
 Organization Type: [National Institutions](#)
 Organization URL: <http://netpreserve.org/>
 Description: The International Internet Preservation Consortium (IIPC) is a membership organization dedicated to improving the tools, standards and best practices of web archiving while promoting international collaboration and the broad access and use of web archives for research and cultural heritage. The Content Development Group leads the IIPC's [collaborative collection projects](#).

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.

Subject Sort By: **Count** | **(A-Z)**

Society & Culture (11)
 Olympics (8)
 Paralympic Games (4)
 Government (2)
 Sports (2)

More ▼

Creator Sort By: **Count** | **(A-Z)**

International Internet Preservation Consortium (12)

Date Sort By: **Count** | **(A-Z)**

December 2013 - February 2014 (1)
 March 2014 (1)

Collector Sort By: **Count** | **(A-Z)**

International Internet Preservation Consortium (12)

Enter search terms here

Collections **Sites** **Search Page Text**

Page 1 of 1 (12 Total Results)

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

2010 Winter Olympics
Archived since: Apr, 2015
Description: A collection of websites related to the 2010 Winter Olympic Games, held in Vancouver, Canada. International Internet Preservation Consortium member institutions contributed suggested websites for inclusion in the collection.
Subject: [Society & Culture](#), [Winter Olympic Games \(21st : 2010 : Vancouver, B.C.\)](#), [Olympics](#)
Creator: [International Internet Preservation Consortium](#)
Collector: [International Internet Preservation Consortium](#)

2012 Summer Olympics
Archived since: Apr, 2015
Description: A collection of websites related to the 2012 Summer Olympic Games, held in London, England. International Internet Preservation Consortium member institutions contributed suggested websites for inclusion in the collection.
Subject: [Society & Culture](#), [Olympic Games \(30th : 2012 : London, England\)](#), [Olympics](#)
Creator: [International Internet Preservation Consortium](#)
Collector: [International Internet Preservation Consortium](#)

The National Archives

BBC Archives

Parliamentary Archives

Public Record Office Northern Ireland

Alan Turing Institute

RESAW

National Library of Ireland

School of Advanced Studies

Thank-you!

E: nicola.bingham@bl.uk

T: @NicolaJBingham
@UKWebArchive

W: www.webarchive.org.uk

“Can you see me mum?”

http://www.webarchive.org.uk/wayback/archive/20100223122017/http://www.oneandother.co.uk/participants/steveplatt

New Tab

UK WEB ARCHIVE

Plinther login | [Plinthers](#) | What's Happening | The project

ONE & OTHER
in partnership with
skyARTS
6 July – 14 October

2400 Places.
34520 Applicants.

skyARTS

CAN YOU SEE ME MUM?

STEVEPLATT
Region: London

Quite nice really

06 Jul | 02PM

NO IMAGE

47:15 24:45/59:49

11 LIKE THIS BIT

This is a live webstream that may contain offensive content. The actions/opinions featured are the participants', and are not endorsed by Artichoke or Sky Arts.

About me

Trafalgar Square is one of my favourite places - part of my London, my England, stuffed full of memories, starting with my first trip to London with my grandma when I was three right through to the present day. What will I do with my hour on the fourth plinth, apart from savour every second of it?

Pledges to watch
Be the first to Pledge your support!

WE'LL WATCH