

Web & Social Media Archiving for Community & Individual Archives: a DPC Briefing Day

Nicola Bingham and Helena Byrne

**Web Archiving
British Library**

December 6th 2018 London

UK Web Archive Overview

- British Library began archiving websites in 2004
- Small scale, selective basis
- Between 2004 and 2013 archived c. 15,000 websites
- Legal Deposit Libraries Act 2003 and LDL(Non-Print Works) Regulations 2013
- Legislative Framework for collecting websites
- UK Legal Deposit Libraries:
 - British Library
 - National Library of Scotland
 - National Library of Wales
 - Bodleian Libraries Oxford
 - Cambridge University Library
 - Trinity College Dublin



UK Web Archive: size

On an annual basis the UK Web Archive acquires:

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

The total collection to date = 470 TB compressed data

" [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]

Non- Print Legal Deposit Regulations

- April 2013 NPLD Regulations
- Enable the Legal Deposit Libraries to archive the UK Web at scale
- 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*]



Non- Print Legal Deposit Regulations 2

IN SCOPE:

- All websites with a .uk domain name (.scot; .cymru etc.)
- UK hosting: check external IP geo-location database for location of server
- Additional, manual checks
 - *UK postal address*
 - *Correspondence with website owner*
 - *Professional judgement*

OUT OF SCOPE:

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

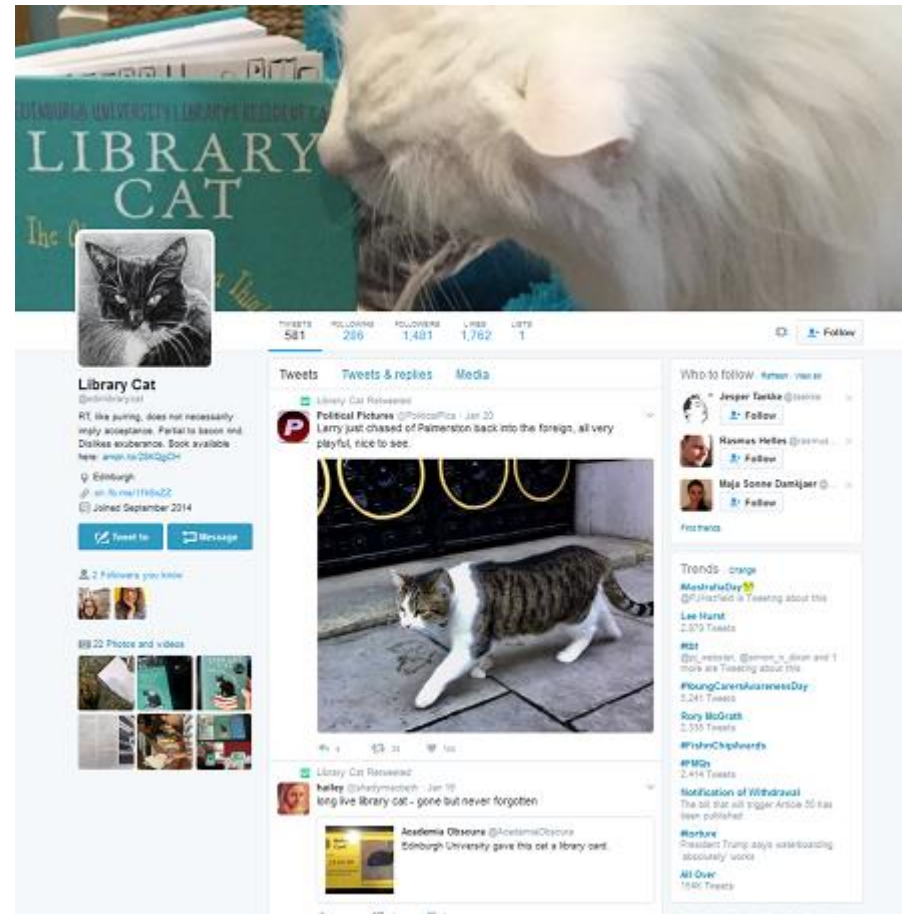
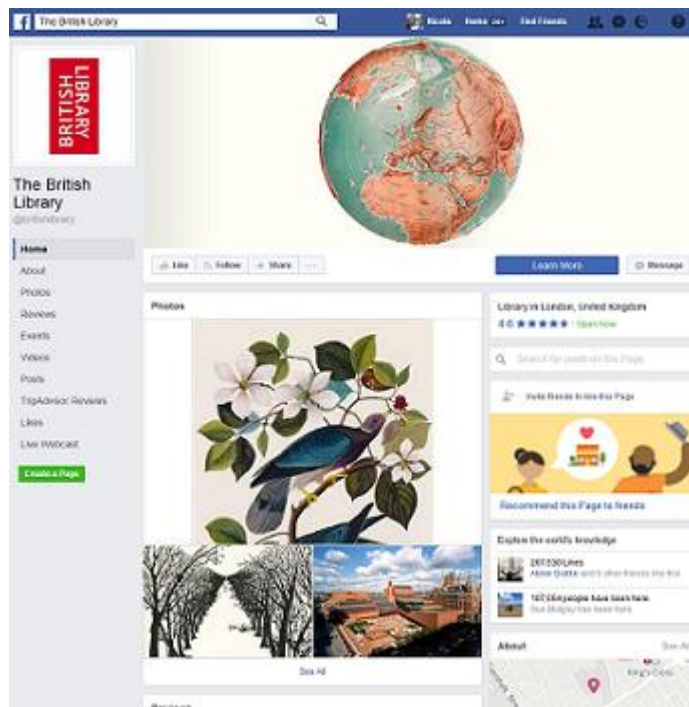
Social Media Platforms

- We can't archive all social media platforms
- Facebook is very locked down
- Youtube – technical and legal restrictions
- Twitter – better results



Social Media

- Collected selectively
- Manually scoped in
- Subject Specialists



Social Media and Personal data

- Social Media important as research dataset.
- Duty to protect individual's privacy.
- No specific and legal “right to be forgotten”
- Legal Deposit Libraries have derogation under the following terms:
 - In order to comply with a legal obligation;
 - For the performance of a task carried out in the public interest;
 - And for archiving purposes.
- Mitigating exposure of personal material:
 - Manual review
 - Access Restrictions
 - Notice and takedown



British Library Reading Room, St. Pancras

Personal Data Checklist General Principles

Do not take down

The material is publicly available.

It was openly published and widely reported.

It was published by the data subject (a responsible adult)

No sensitive information contained.

Temporarily take down and refer

This is the only copy of the material

It was published to a small group or for a limited time period.

Lawfully published by the data subject but as a minor or as a vulnerable adult.

Contains sensitive personal information about the data subject, e.g. racial or ethnic origin, political opinions, religious beliefs, membership of a trade union, physical or mental health or condition, sexual life, criminal offence, court appearance, or financial data.

Take down

n/a

Mistakenly published private data (email, private social network).

Leaked or illegally published without the data subject's consent.

Inherently sensitive data.

Could represent a real threat to the data subject's well-being.

Published/Public data vs Private



mumsnet
by parents for parents



Talk

Advanced search

[Active](#) | [I'm on](#) | [I'm watching](#) | [I started](#) | [Last 15 minutes](#) | [Last hour](#) | [Last Day](#)

Talk » AIBU?

[Start](#) new thread in this topic | [Flip](#) this thread | [Refresh](#) the display

[Add a message](#)

This is page 1 of 1 (This thread has 19 messages.)

To get narked with people spelling 'Paralympics' as 'Paraolympics'? (19 Posts)

Fri 31-Aug-12 11:11:14

[Add message](#) | [Report](#)

I've seen so many posts on FaceBook in which people insist on spelling 'Paralympics' as 'Paraolympics'...or even 'Parolympics' (a kind of bastardisation of the two mis-spellings) and its really grating on me. Almost as much as misused apostrophes!

AIBU? Should I just get over myself?

**Thank-you! ... and
now for some more...**

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture
2. Send out crawl bots

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture
2. Send out crawl bots
3. Download websites into WARC files

How do we ‘archive’ the web?

1. Identify targets (websites) ‘in scope’ (UK) for capture
2. Send out crawl bots
3. Download websites into WARC files
4. Index the collection (so all the words can be searched)

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture
2. Send out crawl bots
3. Download websites into WARC files
4. Index the collection (so all the words can be searched)
5. 'Playback' via a website interface (webarchive.org.uk)

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture
2. Send out crawl bots
3. Download websites into WARC files
4. Index the collection (so all the words can be searched)
5. 'Playback' via a website interface (webarchive.org.uk)
6. Carry out Quality Assurance (QA) on *some* content

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture
2. Send out crawl bots
3. Download websites into WARC files
4. Index the collection (so all the words can be searched)
5. 'Playback' via a website interface (webarchive.org.uk)
6. Carry out Quality Assurance (QA) on *some* content
7. Request open access licence for *some* websites

How do we 'archive' the web?

1. Identify targets (websites) 'in scope' (UK) for capture

Not all UK published websites are on a UK Top Level Domain or hosted in the UK.

[Twitter.com](https://twitter.com)

wordpress.com

wordpress.org

How do we 'archive' the web?

2. Send out crawl bots

We only use Heretrix which is good for large scale crawling but not for getting a high fidelity capture of content that uses a lot of mixed media.

Individual tools like Webrecorder and Social Feed Manager is better for these types of content.

How do we 'archive' the web?

3. Download websites into WARC files

Although we have experimented with Webrecorder and Social Feed Manager the biggest issue we have is ingesting these external files into the archive.

How do we ‘archive’ the web?

5. ‘Playback’ via a website interface (webarchive.org.uk)

***Archived webpages
should look and operate
as they did originally***

Live web

Eventbrite

Search for events

BROWSE EVENTS

HELP

SIGN IN

CREATE EVENT



MAR
08

Upfront and Onside: The Women's Football Conference

by National Football Museum

£85



TICKETS

DESCRIPTION

As part of the Hidden Histories project, funded by the Arts Council, the National Football Museum is delighted to be running and hosting the biggest conference on the history and heritage of women's football. Coinciding with International Women's Day on 8 March, Upfront & Onside features a programme of influential keynote speakers from around the globe focussing on over 150 years of the women's game.

Event Programme 8 March

09:00 – 09:30: Registration Tea & Coffee

09:30 – 09:45: Welcome and Opening Remarks Belinda Monkhouse NFM/ Professor Jean Williams, Professor of Sport and Heritage, University of Wolverhampton

09:45 – 10:45: Keynote Speaker Recovered Memories: Art,

DATE AND TIME

Thu, 8 Mar 2018, 09:00 –
Fri, 9 Mar 2018, 17:15 GMT
[Add to Calendar](#)

LOCATION

National Football Museum
Todd Street
Manchester
M4 3BG
[View Map](#)

REFUND POLICY

Archived web - Open Wayback

External links, forms and search boxes may not function within this archived website.

<https://www.eventbrite.co.uk/e/upfront-and-onside-the-womens-football-conference>

0 captures

2 Feb 18 - 5 Feb 18

JAN

FEB

MAR

2017

2018

2019

Close

Cymraeg

Help

Eventbrite, and certain approved third parties, use functional, analytical and tracking cookies (or similar technologies) to understand your event preferences and provide you with a customised experience. By closing this banner or by continuing to use Eventbrite, you agree. For more information please review our [cookie policy](#).

Eventbrite

Search for events

BROWSE EVENTS

HELP

SIGN IN

CREATE EVENT

MAR 08

Upfront and Onside: The Women's Football Conference

by National Football Museum

£85

TICKETS

DESCRIPTION

As part of the Hidden Histories project, funded by the Arts Council, the National Football Museum is delighted to be running and hosting the biggest conference on the history and heritage of women's football. Coinciding with International Women's Day on 8 March, Upfront & Onside features a programme of influential keynote speakers from around the globe focussing on over 150 years of the women's game.

Event Programme 8 March

DATE AND TIME

Thu, 8 Mar 2018, 09:00 – Fri, 9 Mar 2018, 17:15 GMT

[Add to Calendar](#)

LOCATION

National Football Museum
Todd Street
Manchester

Archived web - Python Wayback

UK WEB
ARCHIVE

Upfront and Onside: The Women's Football Conference Tickets, Thu, 8 Mar 2018 at 09:00 | Eventbrite
2/3/2018, 12:00:57 PM

Back to Calendar
Language: en / c

Eventbrite


Search for events

BROWSE EVENTS

HELP

SIGN IN

CREATE EVENT



MAR
08

**Upfront and Onside:
The Women's Football
Conference**

by National Football Museum

£85

TICKETS

DESCRIPTION

As part of the Hidden Histories project, funded by the Arts Council, the National Football Museum is delighted to be running and hosting the biggest conference on the history and heritage of women's football. Coinciding with International Women's Day on 8 March, Upfront & Onside features a programme of influential keynote speakers from around the globe focussing on over 150 years of the women's game.

Event Programme 8 March

09:00 – 09:30: *Registration Tea & Coffee*

09:30 – 09:45: *Welcome and Opening Remarks* Belinda Monkhouse NFM/ Professor Jean Williams, *Professor of Sport and*

DATE AND TIME

Thu, 8 Mar 2018, 09:00 –
Fri, 9 Mar 2018, 17:15 GMT
[Add to Calendar](#)

LOCATION

National Football Museum
Todd Street
Manchester
M4 3BG
[View Map](#)

Archived web – Current Twitter Display

UKWA Loading... [Back to Calendar](#)
Language: cy / e

[Unmute @FootyCon](#) | [Mute @FootyCon](#) | [Follow @FootyCon](#) | [Following @FootyCon](#) | [Unfollow @FootyCon](#) | [Blocked @FootyCon](#) | [Unblock @FootyCon](#) | [Pending follow request from @FootyCon](#)
[Cancel your follow request to @FootyCon](#)

FootyCon

@FootyCon

The International Football History Conference 7-8 June 2018. Keep the dates free - call for abstracts coming soon!

Joined April 2017
[92 Photos and videos](#) [Photos and videos](#)

Tweets

- Tweets Tweets, current page.
- [Tweets & replies](#)
- [Media](#)

You blocked @FootyCon

Are you sure you want to view these Tweets? Viewing Tweets won't unblock @FootyCon

[Yes, view profile](#)
[Close](#)

FootyCon followed

- [FootyCon](#) Retweeted
[Gary James @GaryJamesWriter](#) Jan 2

[More](#)
 - [Copy link to Tweet](#)
 - [Embed Tweet](#)

Gary James Retweeted The Bertieful South

Archivists? Historians? Fancy a challenge at [@SpursOfficial](#)? Interesting vacancy.....<https://twitter.com/bertiefulsouth/status/948255725486305280>...

Gary James added,

The Bertieful South @BertiefulSouth
[@GaryJamesWriter](#) Any interest to your colleagues? <http://www.tottenhamhotspur.com/jobs/archivist-and-records-manager/> ...
 0 replies 2 retweets 4 likes

[Reply](#)
[Retweet](#) [Retweeted](#)

<https://www.webarchive.org.uk/wayback/en/archive/20180108140418/https://twitter.com/FootyCon/>

How do we 'archive' the web?

6. Carry out Quality Assurance (QA) on *some* content

Limited to just a visual comparison of the archived vs live web.

Use web developer plugins to identify why content is not archived rather than how to get a more complete capture.

The scale is so large we only look at a very small amount of content and this is usually in response to an enquiry.

How do we ‘archive’ the web?

7. Request open access licence for *some* websites

There are separate negotiations going on with large scale publishers in the UK.

Can't always find contact details on a website.

Don't have the resources to look at all the content we archive.

Community Based Websites



The Football Collective

Bringing critical debate to our game

The Collective

Conference ▾

Contribute

Topics ▾

Podcasts

Join the Collective

Films

Youtube

Category: Films



Allison Thompson - Youth Football as a Tool for Cultural (re) Integration

Allison Thompson – Youth Football as a Tool for Cultural (re) Integration

Video of Allison Thompson (International Academy Berlin / Institute Heritage Studies) delivering her presentation at the Football, Politics and Popular ... [More](#)



Alexandra Culvin – New realities for professional women footballers in England

Video of Alexandra Culvin (UCLAN, UK) of her presentation at the Football, Politics and Popular Culture conference, Limerick (2017). Please ... [More](#)



Simon McKerrrell - Kicking metaphors of the body around in the mediation of Self and Other

Simon McKerrrell – Kicking metaphors of the body around in the mediation of Self and Other

Video of Simon McKerrrell (Newcastle University, UK) of his keynote at the Football, Politics and Popular Culture



The Football Collective

Bringing critical debate to our game

The Collective

Conference ▾

Con

Topics ▾

Podcasts

Join

Films

Youtube

Podcasts

Below is the links to The Football Collective podcasts. All members are invited to contribute a podcast based on their expertise, research and experience in football. Below is a list of podcasts to date.

- Podcast with @OnTheBaw by Jennifer Jones hosted on Jen's site [here](#).
- Podcast by Jon Mackenzie hosted on Jons site [here](#)
- Interview with Barry Drust by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Sol Wolfers by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Alex Culvin by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Kieran Maguire by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Jeff McCarthy by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Dr Jim O'Brien by Joshua Dean – [SoundCloud](#) / [iTunes](#)
- Interview with Dougie Brimson by Joshua Dean – [SoundCloud](#) / [iTunes](#)

<https://footballcollective.org.uk/>

Design websites that are archivable?

1. Ensure image, video and audio are NOT coming from somewhere else (Soundcloud, Youtube, Flickr)

Design websites that are archivable?


1. Ensure image, video and audio are NOT coming from somewhere else (Soundcloud, Youtube, Flickr)
2. If you have a database driven site, include a sitemap

Design websites that are archivable?

- 1. Ensure image, video and audio are NOT coming from somewhere else (Soundcloud, Youtube, Flickr)**
- 2. If you have a database driven site, include a sitemap**
- 3. Use robots.txt to prevent access to areas of the site which may cause problems if crawled e.g. databases, including online catalogues; "shopping baskets", etc.**

<http://blogs.bl.uk/webarchive/2012/09/how-to-make-websites-more-archivable.html>

www.archiveready.com

 **ArchiveReady**
website archivability evaluation tool

HelpFAQAPI

Checking website: <https://www.bl.uk/> All messages

SummaryHTML and CSS 9HTTP 4Media 48Sitemaps 3

Overall Rating
84%
One page printable HTML
EARL XML results

Web Archivability Facet	Rating
Accessibility	59%
Cohesion	87%
Metadata	100%
Standards Compliance	92%

About the method


Web archivability can be measured from several perspectives. Here, we have called these perspectives Archivability Facets. Their selection and calculation is based on information gathered from the target website combined with an evaluation of the website's compliance with recognised practices in digital curation (e.g. using adopted standards, validating formats, and assigning metadata). For more information, please check the following publications:

Banos V., Manolopoulos Y., Web Content Management Systems Archivability, ADBIS 2015, BIB.

Banos V., Manolopoulos Y.: A quantitative approach to evaluate Website Archivability using the CLEAR+ method, International Journal on Digital Libraries, 2015, Springer Link · BIB.

Banos V., Kim Y., Ross S., Manolopoulos Y.: CLEAR: a credible method to evaluate website archivability, IPRES 2013, PDF · BIB.

Banner in your website

You can also use the following html code to add this banner to your website: 

```
<a href="http://archiveready.com/check?url=https://www.bl.uk/" title="Website Archivability Testing"></a>
```

Linking to this result

If you would like to create a link to this validation result to make it easier to revalidate this page in the future or to allow others to validate your page, the URI is <http://archiveready.com/check?url=https://www.bl.uk/>.

History

Date	Website Archivability	Acc	Coh	Met	Sta
2017-10-30 13:21:27	52%	16%	80%	100%	11%

Is your website Archive Ready?
 Check now »

HomeHelpFAQAPI

ArchiveReady is a project by Vangelis Banos (© 2012-2017).

www.bl.uk

33

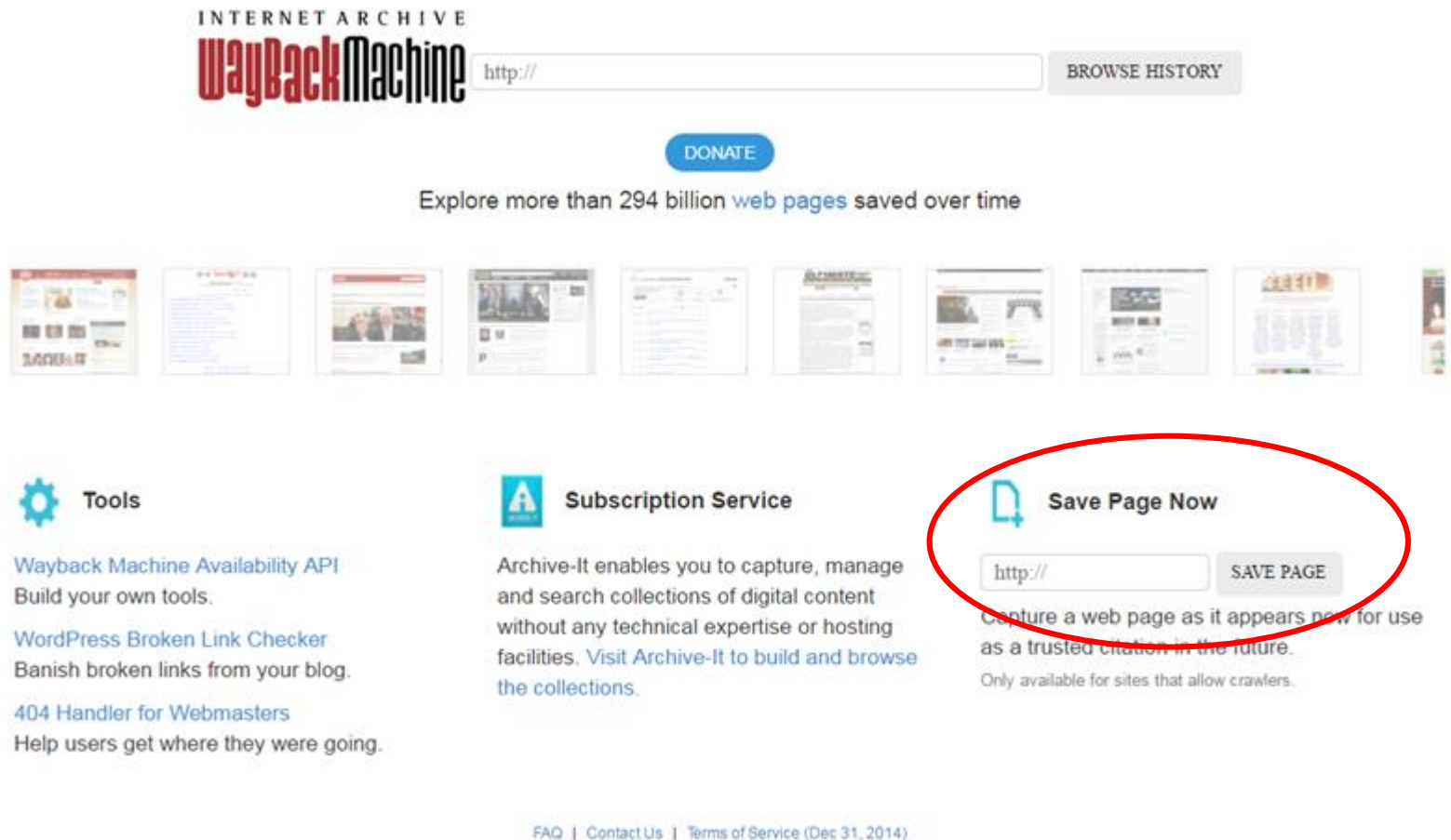
You Can Get Involved

Save a UK website - Nominate now!



beta.webarchive.org.uk/en/ukwa/info/nominate

Save content from outside the UK














The image shows the Wayback Machine homepage. At the top, it says "INTERNET ARCHIVE" and "WayBackMachine". There is a search bar with "http://" and a "BROWSE HISTORY" button. Below that is a "DONATE" button. A banner says "Explore more than 294 billion web pages saved over time". A row of ten small thumbnail images of various websites is shown. Below this, there are three main sections: "Tools", "Subscription Service", and "Save Page Now". The "Save Page Now" section is circled in red. It contains a search bar with "http://", a "SAVE PAGE" button, and text: "Capture a web page as it appears now for use as a trusted citation in the future. Only available for sites that allow crawlers."

INTERNET ARCHIVE
WayBackMachine

http://

Explore more than 294 billion [web pages](#) saved over time


         

 **Tools**


[Wayback Machine Availability API](#)
Build your own tools.

[WordPress Broken Link Checker](#)
Banish broken links from your blog.

[404 Handler for Webmasters](#)
Help users get where they were going.

 **Subscription Service**

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit [Archive-It](#) to build and browse the collections.

 **Save Page Now**

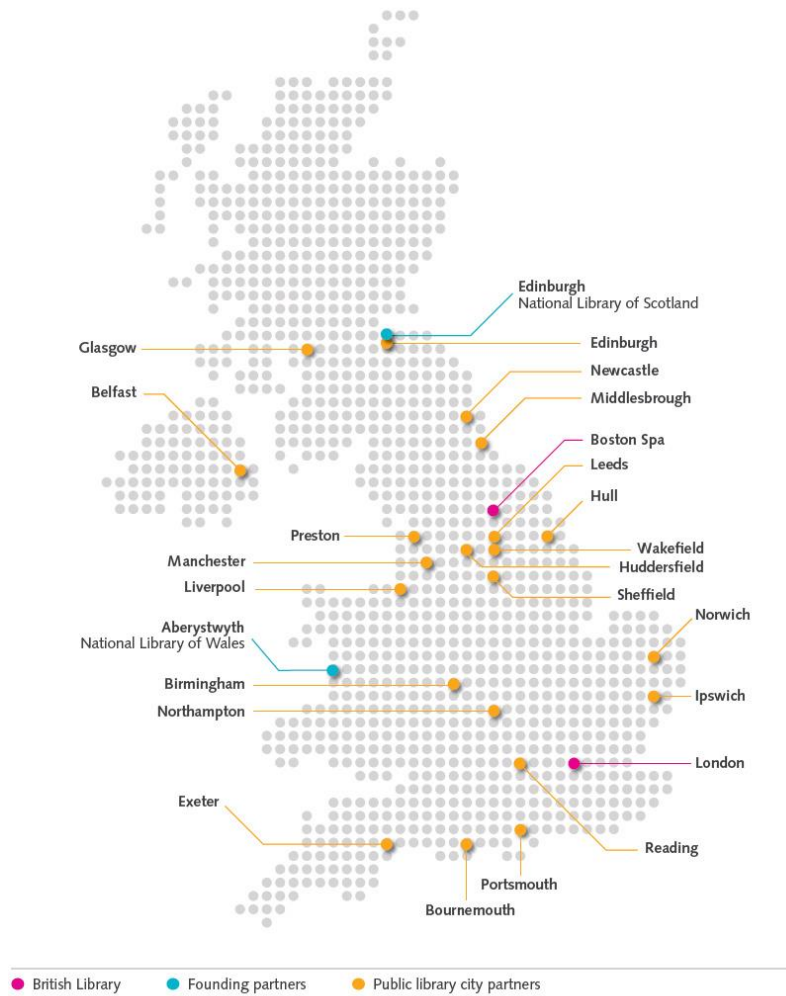
http://

Capture a web page as it appears now for use as a trusted citation in the future.
Only available for sites that allow crawlers.

[FAQ](#) | [Contact Us](#) | [Terms of Service \(Dec 31, 2014\)](#)

<https://archive.org/web/>

Living Knowledge Network



Your web archive needs YOU!

- Save a website! - Nominate

Your web archive needs YOU!

- Save a website! - Nominate
- Create a Collection – Get in touch

Your web archive needs YOU!

- Save a website! - Nominate
- Create a Collection – Get in touch
- Ask **five friends** if they know a UK website that should be saved

Useful Links

webarchive.org.uk

webarchive.org.uk/blog

webarchive.org.uk/videos

webarchive.org.uk/shine

data.webarchive.org.uk/opendata

