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

- 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

<table>
<thead>
<tr>
<th>Domain</th>
<th>TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>477.64 TB</td>
</tr>
<tr>
<td>Frequent</td>
<td>255.24 TB</td>
</tr>
<tr>
<td>Total</td>
<td>732.88 TB</td>
</tr>
<tr>
<td>Webrecorder</td>
<td>1.63 TB</td>
</tr>
</tbody>
</table>

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

There are (many) technical limitations

• Database driven sites
• Programming Scripts
• Plug-Ins
• Proprietary file formats
• Blockers
  – Robots.txt
  – Access denied
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?
Collaboration IIPC

- 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?”

This is a live webstream that may contain offensive content. The actions, opinions featured are the participants’, and are not endorsed by Anticlock or SkyArts.

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?