10 Years of archiving the UK Web

Helen Hockx-Yu
Head of Web Archiving, British Library
Web Archiving @ the British Library

- **Explore**
  - Domain.UK project
  - 100 websites
  - No public access
  - 2001-2002

- **Collaborate**
  - Web Archiving Programme
  - Led UK Web Archiving Consortium
  - The UK Web Archive
  - 2003-2007

- **Build capacity**
  - Improvement and new capabilities
  - Curatorial expertise
  - Technical know-how
  - Researcher engagement
  - 2008-2011

- **BAU**
  - Business as Usual
  - Implement non-print Legal Deposit since April 2013
  - 2011

Collect UK digital heritage and provide continued access to archived web resources
The UK Web Archiving Consortium
6th April 2013...

- Legal Deposit Libraries (Non-Print Works) Regulations 2013
- Extension of existing legal framework
- Systematic collection of UK’s published output for heritage & preservation
- By 6 UK Legal Deposit Libraries
The Digital Library System

- 4 Nodes (Complete Copies)
  - British Library, St. Pancras
  - British Library, Boston Spa
  - National Library of Wales
  - National Library of Scotland

- Additional Access Points
  - Bodleian Library, Oxford
  - Cambridge University Library
  - Trinity College Library, Dublin
The UK Web Domain

- Over 10 million .uk registered domain

- UK organisations also use non .uk domain names (eg .com or .org)
  - Exact scale unknown

- Non-print Legal Deposit applies to
  - the open (freely available) web: .uk
  - other UK-published (non .uk) websites, such as .com, .org...
    - made available to the public by a person or an organisation and the activities relating to the creation or the publication of the work take place within the United Kingdom
Collecting strategy

Domain Crawl

Events
- Domain crawl: Broad crawl of UK domain
- Once or twice a year

Key sites
- Events & key sites and news:
  - Events of UK interest
  - High value, high impact sites
  - National & regional news

News
- Special Collection:
  - Focused, thematic collections
  - Support priority subjects
Open UK Web Archive

- Based on websites owners’ permissions
- 24/7 access
- ~68,000 point-in-time snapshots of over 15,000 selected UK websites
- 21TB
- Continues to grow
- Going forward: showcase the Legal Deposit content
Legal Deposit UK Web Archive

- Annual UK Domain Crawl
  - 2013: 31TB, 1.6 billion URLs, 4 million hosts
  - 2014: 57TB including 2.5 non .uk domains
- Events, news sites, key sites
  - ~15TB
- Discoverable through British Library’s online catalogue
- Full-text searchable in the reading rooms
“Hockx-Yu is now on the front line of the most ambitious expansion of the British Library’s archiving capability in more than 300 years. At the stroke of midnight on April 5, 2013, legislation known as the Legal Deposit regulations came into force, charging the library with capturing the contents of the entire U.K. web domain—every site carrying the .uk suffix—preserving the material and making it publicly accessible.”

Jacobson (2014)
Web archiving: the global picture

- Survey of web archiving initiatives (Daniel Gomes et al 2010)
  - 42 web archiving initiatives across 26 countries since 1996
  - 6.6PB archived web resources

- International Internet Preservation Consortium (IIPC) - from 12 founding members in 2003 to nearly 50 in 2015
  - Most members are national memory institutions

- How much of the Web is Archived (Scott Ainsworth et al, 2012)

<table>
<thead>
<tr>
<th>Percentage archived</th>
<th># of copies in public archive</th>
</tr>
</thead>
<tbody>
<tr>
<td>35% -90%</td>
<td>At least one</td>
</tr>
<tr>
<td>17-49%</td>
<td>2-5</td>
</tr>
<tr>
<td>1%-8%</td>
<td>6-10</td>
</tr>
<tr>
<td>8%-63%</td>
<td>&gt;10</td>
</tr>
</tbody>
</table>

- BL and CDNL (association of chief executives of national libraries) survey: 30 nations have legislations by 2012 enabling web harvesting
Types of web archives

- Global: Internet Archive’s Wayback Machine

- National: UK, Denmark, France, Finland and many other countries

- Sub-set of a national domain: eg UK Government Web Archive, Canadian Government Web Archive

- Topical collections: eg Human rights

- Collaborative: Archive–It service by Internet Archive

- Corporate: a single organisation’s web estate, eg UK Parliament Web Archive
Reflections / observations

- High concentration of use of technology and practice
- Modern browsers can render virtually any webpage. But the same pages often defeat our purpose-built crawlers
- National approach to archiving the global web
- Does size matter?
- Role of Legal Deposit
- Library/Archive discourse
- What is driving web archiving?
Access and use of web archives

- Completeness versus openness
  - Many countries don’t have Legal Deposit
  - Legal Deposit national collections have restricted access

- Not enough versus too much
  - Pre-selected or defined collections not relevant to all researchers; difficulty in finding relevant content in large scale web archive

- Arbitrary (national) boundaries often irrelevant to research question

- (Single) envisaged use case based on browsing of individual websites page by page
An archived website
Collections
Subjects
Legal Deposit UK Web Archive

- Full-text search, results can be refined by facets
  - Content type (html, pdf…)
  - Crawl year
  - (individual) Domain (bbc.co.uk; newsforscotland.com…)
  - Domain suffix (.co.uk; .org.uk; .ac.uk; .com)
  - Author (based on metadata extracted from web pages)
  - …

- Still has “collections” – selected and curated by curators, usually crawled within a fixed period of time, relating to an event

- Subjects information – added to selected sites only – included in the index
Issues of “document-centric” approach

- Rise of “digital scholarship”, taking advantage of the possibilities offered by technology
- Primacy of “text” as object of study no longer exists
- “Paratexts” play a crucial role in textual coherence of a website: header & footer drop-down menu, site map, breadcrumb, etc. (Brügger, 2010)
- “Distant reading” focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems (Moretti, 2000)
“Resource Not in Archive”

- Common error message but appears for different reasons

- Intended boundary
  - No permission for linked content
  - Not allow by robots.txt
  - Edge of a archive
  - Data limitation

- Technical limitations, e.g. dynamic content the crawler could not collect

- In UK Web Archive: avoid dead end in navigation
  - Search
  - Link to live web
  - Find archived copies elsewhere
There is so much more…

- Statistical overview, scale and distribution of a web national domain
- Evolution, pattern of change over time, eg domain names
- Space: geo location, postcodes
- Type of content, eg file format, language
- Correlation, eg between certain term and historical event
- Structure, linked entities and networks
- Other web archives
JISC UK Web Domain dataset (1996-2013)

- Copies of UK websites extracted from the Internet Archives collection
  - 57TB, 4 billion URLs

- Research agreement between JISC and IA, upholding IA’s Terms of Use
  - Access via IA’s Wayback Machine
  - Allows replication / extraction of secondary datasets: Format profile, Geoindex, Host Link Graph

- BL hosts the dataset on behalf of JISC

- Serves as research test bed and used by various research projects
  - JISC funded: Analytical Access to the Domain Dark Archive (AADDA); Big data for political science
  - AHRC funded: Big UK Domain Data for Arts and Humanities
  - ERC funded: Alexandria
The UK Web in 2013

Largest UK Hosts (in GB)

Number of Hosts

Total Data Volume

# Hosts

Number of URLs
Evolution of the UK web (2004 - 2013)
Postcode-based access
HTML version analysis

The chart illustrates the percentage of HTML resources over time, with different versions of HTML. The following versions are shown:

- HTML 2.0
- HTML 3.2
- HTML 4.0
- HTML 4.01
- XHTML 1.0

The chart shows a decline in the use of older versions of HTML and an increase in the use of newer versions from 1996 to 2010.
Image Format Analysis

The graph shows the percentage of crawls for different image formats over the years from 1996 to 2010. The formats include JPEG, GIF, PNG, Icon, XBM, and TIFF. The y-axis represents the percentage of crawls, while the x-axis represents the years. The data indicates a decline in the percentage of crawls for all formats except PNG, which shows an increase in the later years.
Using N-gram for scholarly research

- Courtesy of Dr Peter Webster, Institute of Historical Research, University of London
Exploring links

Figure 5: Network diagram of hyperlinks between universities. Different colors indicate different university affiliations.
Visualising links (to and from bl.uk 1996)
Visualising links (to and from bl.uk 2010)
How was the UK web linked in 1996?

- By Rainer Simon using UK Host-Level Link Graph (1996-2010) dataset.
- Based on the 1996 portion: 58,842 hosts (nodes); 184,433 host-to-host links (edges)
- UK web as part of the global web
- Scalability issues with large dataset over time
Mementos service (http://www.dcc.ac.uk)
Collaboration with researchers

- Building collections
  - Researchers’ involvement in scoping collections, selecting and describing websites
  - Creation of specific, (narrow) topical collections

- Formulating research question
  - Brain-storm sessions, workshops, discussion, surveys etc.
  - Lack of awareness & baseline knowledge
  - Challenging: you don’t know what you don’t know
  - Analytical access: web archive as dataset
Co-development

- Institute of Historical Research project: Analytical Access to the Domain Dark Archive (AADDA)
  - Experimental user interface

- Big UK Domain Data for Arts and Humanities (2014-2015)
  - Institution of Historical Research, Oxford Internet Institute, British Library and Aarhus University
  - Develop theoretical and methodological framework for the study of web archives
  - Build on ADDAA: researchers and the BL co-produce access tools
  - Study of the history of UK web space from 1996 to 2013
  - Research use cases

- Changing the way we archive the web
Welcome to our 11 bursary holders

One of the main aims of the project is to involve arts and humanities researchers in the development of tools for analysing web archives, thereby ensuring that those tools meet real rather than perceived researcher needs. We recently ran an open competition inviting researchers to submit proposals across a range of disciplines which focus on the archived web, and have selected 11 from a tremendously strong and varied set of applications. The topics that will be studied over the next eight months are:

- Rowan Aust – Tracing notions of heritage
- Rona Cran – Beat literature in the contemporary imagination
- Richard Deswarte – Revealing British Eurosceptism in the UK web domain and archive
- Chris Fryer – The UK Parliament Web Archive
- Saskia Huc-Heapher – An ethnomesiotic study of London French habitus as displayed in blogs
- Alison Kay – Capture, commemoration and the citizen-historian: Digital Shoebox archives relating to P.O.W.s in the Second World War
- Gareth Millward – Digital barriers and the accessible web: disabled people, information and the internet
- Marta Musso – A history of the online presence of UK companies
- Harry Raffal – The Ministry of Defence’s online development and strategy for recruitment between 1996 and 2013
- Lorna Richardson – Public archaeology: a digital perspective
- Helen Taylor – Do online networks exist for the poetry community?
Shine

- Query building
- Corpus formation and handling
- Annotation and curation
- In-corpus analysis
- Whole-dataset analysis
Scholarly use of web archives

http://netpreserve.org/web-archiving/videos
## Requirements for web archives

<table>
<thead>
<tr>
<th>Scholarly requirements</th>
<th>Requirements for web archives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>No access restriction, available online</td>
</tr>
<tr>
<td>Text, paratext and context</td>
<td>- Access to text and any contextual information at various granularity.</td>
</tr>
<tr>
<td></td>
<td>- Access to secondary datasets</td>
</tr>
<tr>
<td>Persistence and citability</td>
<td>- Persistent identifiers</td>
</tr>
<tr>
<td></td>
<td>- Standards of citing archived websites</td>
</tr>
<tr>
<td></td>
<td>- Integration with bibliographical management tools (eg Zotero)</td>
</tr>
<tr>
<td>Formation of research corpus</td>
<td>- Archiving of research corpora on demand</td>
</tr>
<tr>
<td></td>
<td>- Means to mix, match and reassemble corpora</td>
</tr>
<tr>
<td>Quality</td>
<td>- Represents as much as possible the live website in completeness, intellectual content, behaviour and look and feel</td>
</tr>
<tr>
<td></td>
<td>- Information on what is “missing” to help interpret incomplete sources</td>
</tr>
<tr>
<td>Close and distant reading</td>
<td>Multiple access methods including tools for data analytics and visualisations</td>
</tr>
<tr>
<td>Boundary &amp; format-independent</td>
<td>- Interlinked web archives</td>
</tr>
<tr>
<td></td>
<td>- integration with other digital and printed sources</td>
</tr>
</tbody>
</table>
IDCC 2025

- The UK Web Archive
  - Single search for all collections
  - Base-line knowledge self-explanatory, text and much contextual information
  - Link to many more web archives
  - Inactive websites highlighted
  - Open datasets / content

- “Macroscope” of the UK web history
  - “a single data point, .. both visualised at scale in the context of a billion other data points, and drilled down to its smallest compass”

- Flexibility to reconstruct context, big or small

- Seamlessly linked web archives

- Crawlers as capable as browsers

- Independent use of web archives
Java Developer for Web Archiving

Closing date Midnight 15th February

Please help spread the word or Apply!

Contact: Helen.hockxyu@bl.uk
Twitter: @ukwebarchive
References


