Negotiating the archives of UK web space

Jane Winters, Professor of Digital Humanities, School of Advanced Study, University of London
Workshop on National Webs, Aarhus, 8-9 December 2016
Facts and figures I

• Jisc historical dataset 1996 to 6 April 2013
  – 3,520,628,647 distinct records
  – 65 terabytes

• 2014 domain crawl (.uk)
  – 56TB data
  – 2.5 billion webpages and other assets (including 4.7GB of viruses)
Facts and figures II

• UK Parliament Web Archive
  – Three snapshots per year covering 30 sites (37 sites in the archive in total)
  – 4.8TB data

• UK Government Web Archive
  – 3,000+ websites,
  – Twitter (65,000 tweets) and video archives
Internal inconsistencies

• UKGWA consists of data provided by IA 2003-4 (plus back catalogue to 1996); and by the Internet Memory Foundation from 2005 onwards (further complicated by membership of UKWAC)
• The BL annual domain crawl has failed differently each time it has run
• The ‘break’ between IA and nationally archived content
UK PARLIAMENT WEB ARCHIVE
This collection of archived web content was selected for preservation by the Parliamentary Archives.

5,277,655,813,324 bytes archived

Allowances
http://mpsallowances.parliament.uk/mpslordsandoffice
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.

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).
Welcome

Welcome to SHINE, our prototype historical search engine.

About SHINE

This tool has been developed as part of the Big UK Data Arts and Humanities project funded by the AHRC. Read more about the project on our blog.

The data was acquired by JISC from the Internet Archive (IA) and includes all UK websites in the IA web collection crawled between around 1996 until April 2013 (you can read more details about this dataset here). The collection comprises over 3.5 billion items (urls, images and other documents) and has been full-text indexed by the UK Web Archive. Every word of every website in the collection can be searched for and analysed.

What can I do?

Trend analysis

Use ‘trends’ to analyse the number of pages a word or phrase appears in the collection over a given period (within 1996-2013). Comparisons can be drawn by adding several words or phrases separated by a comma. E.g. cat, dog, goldfish
Warning! This is a research prototype for a web archive search service, and may be taken down at any time.

Click on a point on the graph to show a random sample of the matching records from that year...
Click on a point on the graph to show a random sample of the matching records from that year...
data.bl.uk

UK Web Archive

There are millions of UK websites. They are constantly changing and even disappearing. Often they contain information that is only available online. Responding to the challenge of a potential "digital black hole" the UK Web Archive is there to safeguard as many of these websites as practical. Its purpose is to give permanent online access to key UK websites for future generations.

The UK Web Archive contains websites that publish research, that reflect the diversity of lives, interests and activities throughout the UK, and demonstrate web innovation. This includes "grey literature" sites; those that carry briefings, reports, policy statements, and other ephemeral but significant forms of information.

Because websites are revisited and snapshots ("instances") taken at regular intervals, readers can see how a website evolves over time. The archive is free to view, accessed directly from the Web itself and, since archiving began in 2004, has collected thousands of websites.

Since April 2013 the British Library has begun to archive the whole of the UK web domain, under the terms of the Non-Print Legal Deposit Regulations 2013. This is in partnership with the other five legal deposit libraries for the UK: the National Library of Scotland, the National Library of Wales, Cambridge University Library, the Bodleian Library in Oxford, and the Library of Trinity College Dublin. Access to the whole legal deposit web archive will be possible only on premises controlled by one of the UK legal deposit libraries. However, the UK Web Archive team will seek additional permissions from the owners of a small selected group of sites for offsite access through this site.


The dataset is a format profile, summarising media type (MIME-type) data formats contained within all of the HTTP 200 OK responses in the 1996 - 2010 tranche of the JISC UK Web Domain Dataset.

DOI: 10.5255/skwa.2/6/2011

JISC UK Web Domain Dataset GeIndex. 1996 - 2010. TSV.

The dataset comprises ~2.5 billion 200 OK responses in the 1996 - 2010 tranche of the JISC UK Web Domain Dataset which have been scanned for geographic references - specifically postcodes. This set of postcode citations, found at particular URLs and crawled at particular times, forms an historical geindex of the UK web.
National Library of Wales, Cambridge University Library, the Bodleian Library in Oxford, and the Library of Trinity College Dublin. Access to the whole legal deposit web archive will be possible only on premises controlled by one of the six legal deposit libraries. However, the UK Web Archive team will seek additional permissions from the owners of a small selected group of sites for offline access through this site.


The dataset is a format profile, summarising media type (MIME type) data formats contained within all of the HTTP 200 OK responses in the 1996 - 2010 tranche of the JISC UK Web Domain Dataset.

DOI: 10.5255/ukwa.ds.2/fmt/1

JISC UK Web Domain Dataset GeoIndex. 1996 - 2010. TSV.

The dataset comprises ~2.5 billion 200 OK responses in the 1996 - 2010 tranche of the JISC UK Web Domain Dataset which have been scanned for geographic references - specifically postcodes. This set of postcode citations, found at particular URLs and crawled at particular times, forms an historical geoindex of the UK web.

DOI: 10.5255/ukwa.ds.2/geo/1

JISC UK Web Domain Dataset Host Link Graph. 1996 - 2010. TSV.

The dataset comprises ~2.5 billion 200 OK responses from the 1996 - 2010 tranche of the JISC UK Web Domain Dataset which have been scanned for hyperlinks. For each link, UKWA extracts the host that the link targets, and uses this to build up a picture of which hosts have linked to which other hosts, over time.

DOI: 10.5255/ukwa.ds.2/host_linkage/1

JISC UK Web Domain Dataset Crawled URL Index. 1996 - 2013. CDX.

The dataset comprises original compound index (CDX) files that have been re-assembled into 18 separate CDX files for each year of crawling activity represented (1996 - 2013). Please note that the individual CDX files are not sorted. In order to enable access to web archives, UKWA uses CDX files to act as indexes so that it is possible to look up which ARC or WARC files contain which URLs and responses.

DOI: 10.5255/ukwa.ds.2/cdx/1

UK Selective Web Archive Classification Dataset. 1996 - 2010. TSV.

The dataset comprises a manually curated selective archive produced by UKWA which includes the classification of sites into a two-tiered subject hierarchy.

DOI: 10.5255/ukwa.ds.1/classification/1
Comparison of annual datasets, 1996-1998

- Total unique URLs
- Total applications
- Total text
- Total images
- Total audio
- Total video
- Total files

Number of format types, 1996-1997

- Text types
- Image types
- Application types
- Video types
- File types

Number of format types: 1996 - 1997
### Monthly Registration Statistics .uk Domain Names (2008)

<table>
<thead>
<tr>
<th>Month</th>
<th>.co.uk</th>
<th>.me.uk</th>
<th>.org.uk</th>
<th>.ltd.uk</th>
<th>.plc.uk</th>
<th>.net.uk</th>
<th>.sch.uk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>138169</td>
<td>1973</td>
<td>8265</td>
<td>152</td>
<td>3</td>
<td>1</td>
<td>29</td>
<td>148592</td>
</tr>
<tr>
<td>February</td>
<td>133816</td>
<td>1929</td>
<td>8318</td>
<td>142</td>
<td>4</td>
<td>2</td>
<td>29</td>
<td>144240</td>
</tr>
<tr>
<td>March</td>
<td>134708</td>
<td>1903</td>
<td>8404</td>
<td>141</td>
<td>2</td>
<td>1</td>
<td>20</td>
<td>145259</td>
</tr>
<tr>
<td>April</td>
<td>135597</td>
<td>2190</td>
<td>8437</td>
<td>160</td>
<td>1</td>
<td>5</td>
<td>26</td>
<td>146424</td>
</tr>
<tr>
<td>May</td>
<td>126697</td>
<td>1934</td>
<td>7805</td>
<td>125</td>
<td>3</td>
<td>1</td>
<td>38</td>
<td>136503</td>
</tr>
<tr>
<td>June</td>
<td>131424</td>
<td>1952</td>
<td>7756</td>
<td>146</td>
<td>3</td>
<td>4</td>
<td>61</td>
<td>141345</td>
</tr>
<tr>
<td>July</td>
<td>135136</td>
<td>2277</td>
<td>8031</td>
<td>125</td>
<td>6</td>
<td>1</td>
<td>46</td>
<td>145622</td>
</tr>
<tr>
<td>August</td>
<td>124963</td>
<td>2030</td>
<td>7364</td>
<td>120</td>
<td>2</td>
<td>2</td>
<td>26</td>
<td>134515</td>
</tr>
<tr>
<td>September</td>
<td>129834</td>
<td>1958</td>
<td>7901</td>
<td>122</td>
<td>2</td>
<td>1</td>
<td>53</td>
<td>139871</td>
</tr>
<tr>
<td>October</td>
<td>139813</td>
<td>1946</td>
<td>8109</td>
<td>122</td>
<td>2</td>
<td>0</td>
<td>35</td>
<td>143344</td>
</tr>
</tbody>
</table>
Breakdown of domain name registrations, 2000 (Nominet)

- .co.uk: 1,575,655
- .org.uk: 108,711
- .ltd.uk: 4,626
- .plc.uk: 265
- .net.uk: 42
- .sch.uk: 8,830
Acknowledgements

• BUDDAH project team – Jonathan Blaney, Niels Brügger, Josh Cowls, Helen Hockx-Yu, Andrew Jackson, Eric Meyer, Ralph Schroeder, Jason Webber, Peter Webster

• Bursary holders – Rowan Aust, Rona Cran, Richard Deswarte, Saskia Huc-Hepher, Alison Kay, Gareth Millward, Marta Musso, Harry Raffal, Lorna Richardson, Helen Taylor