Ancient History of the UK Web

Josh Cowls, Scott A. Hale, Helen Margetts, Eric T. Meyer, Ralph Schroeder, Taha Yasseri

With support by and thanks to Ning Wang and Adham Tamer
Past Web Archive Activities at OII

  – OII, Internet Archive, Hanzo Archives

• 2010. Researcher Engagement with Web Archives (Jisc funded)
  – OII, VKS

• 2011. Using Web Archives: A Futures Perspective (IIPC funded)
  – OII
Recent Web Archive Activities at OII

• 2013-2015: Jisc Big Data project (Jisc funded)
  – OII, British Library
  – Prepare and release hyperlink corpus

• 2014-2015: Big UK Domain Data for the Arts and Humanities (AHRC funded)
  – IHR, OII, British Library
  – Supporting researchers in Arts & Humanities to use web archive data
  – Producing edited book of empirical studies concerning the history of the UK web

• First paper from these combined projects
Big Data:
Demonstrating the Value of the UK Web Domain Dataset for Social Science Research

This project aims to enhance JISC's UK Web Domain archive, a 30 TB archive of the .uk country-code top level domain collected from 1996 to 2010. It will extract link graphs from the data and disseminate social science research using the collection.
Taming a mammoth:
Web Archive Dataset Preparation

30 TB compressed data

6.2TB metadata and links

2.5 TB temporal links
30 TB compressed data in (w)arc format

- Approx. 4.5 million files
- Mix of binary and plain text payloads along with header data
- Two formats: old arc and newer warc

Housed at the BL, access restrictions
Extract meta-data and links (wat format)

- Approx. 4.5 million files
- 6.2TB on disk compressed
- Housed at OII
- Structured JSON
- Different formats for arc/warcs
"Title": "Exterior Corbels"
Plain text lists

Build own ad-hawk Hadoop cluster, fix incompatibilities, divide into smaller batches
  – Build plain text lists of pages and hyperlinks
  – Remove error page (e.g., 404 Not Found)
  – Remove pages not in .uk
  – Standardize dates (many formats)
  – Standardize hyperlinks (trailing /, etc.)
  – Fix/remove tons of invalid hyperlinks (whitespace, invalid characters, etc.)

Load results into Apache Hive (2.5 TB)
<table>
<thead>
<tr>
<th>Source</th>
<th>Destination</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://octopus.well.ox.ac.uk:80/">http://octopus.well.ox.ac.uk:80/</a></td>
<td><a href="http://octopus.well.ox.ac.uk:80/links.html">http://octopus.well.ox.ac.uk:80/links.html</a></td>
<td>1032758438</td>
</tr>
<tr>
<td></td>
<td>Links</td>
<td></td>
</tr>
<tr>
<td><a href="http://octopus.well.ox.ac.uk:80/">http://octopus.well.ox.ac.uk:80/</a></td>
<td><a href="http://octopus.well.ox.ac.uk:80/projects.html">http://octopus.well.ox.ac.uk:80/projects.html</a></td>
<td>1001793436</td>
</tr>
<tr>
<td></td>
<td>Projects</td>
<td></td>
</tr>
<tr>
<td><a href="http://octopus.well.ox.ac.uk:80/">http://octopus.well.ox.ac.uk:80/</a></td>
<td><a href="http://debian.org/">http://debian.org/</a></td>
<td>1075794060</td>
</tr>
<tr>
<td></td>
<td>Debian/GNU</td>
<td></td>
</tr>
</tbody>
</table>
Overall Statistics

Third-level domains:

e.g. ox.ac.uk
Relative size of second-level-domains
Number of links within SLD per node
Cross-domain links (2010)

Absolute

Normalized to target size
Case of ac.uk

121 UK universities websites and links

1) League table ranking
2) Group affiliation
3) Geographical location

Mapping the UK Webspace: Fifteen Years of British Universities on the Web
Group Affiliations

- No affiliation
- Million+ Group
- Russell Group
- University Alliance
- Cathedrals Group
- 1994 Group
League table ranking

Spearman's rank correlation coefficient

In-Strength Rank

University Rank
Geography

2000  2005  2010

Colour ~ intensity
Gravity Law

\[ \sigma_{ij} = \frac{s_{ij}}{s_{i}^{\text{out}} s_{j}^{\text{in}}} \]

\[ s_{ij} = \frac{s_{i}^{\text{out}} s_{j}^{\text{in}}}{r^{0.28}} \]

Distance (km)

\[ a = 0.28 \pm 0.02 \]
Big UK Domain Data for the Arts and Humanities

Primary aim: developing a methodological and theoretical framework within which to study over 15 years of UK domain data – with lessons for the future study of web archives more generally
Big UK Domain Data for the Arts and Humanities

The dataset:

– Crawled from 1996 – 2013
– Approximately 65 TB, billions of words
– Building interface to allow search by retrieval date, target domain of links, sentiment
– Allow qualitative and quantitative analysis – and iteration between multiple research techniques
Big UK Domain Data for the Arts and Humanities

Key outputs:

– Ten bursary projects using web archive data to investigate a broad range of topics, for example...
  • Armed services recruitment online
  • The accessibility of the web for disabled users
  • Online discussions of ‘Beat’ poetry

– An edited book of empirical studies concerning the history of the UK web, featuring chapters on, for example...
  • Constitutional and institutional change in UK government
  • The BBC’s online presence
  • The ‘web of faith’ online
Next

- Studies underway at OII, BL, IHR
- Book and articles
  - Study overall growth of .uk
  - Case study of .gov.uk
  - Study of media and select committee visibility
- Releasing data open source