Web Archives in the Eyes of Scholars

Helen Hockx-Yu

Head of Web Archiving

British Library

18 April 2013

Web Archives

COVERAGE AND USAGE

How much of the web is archived?

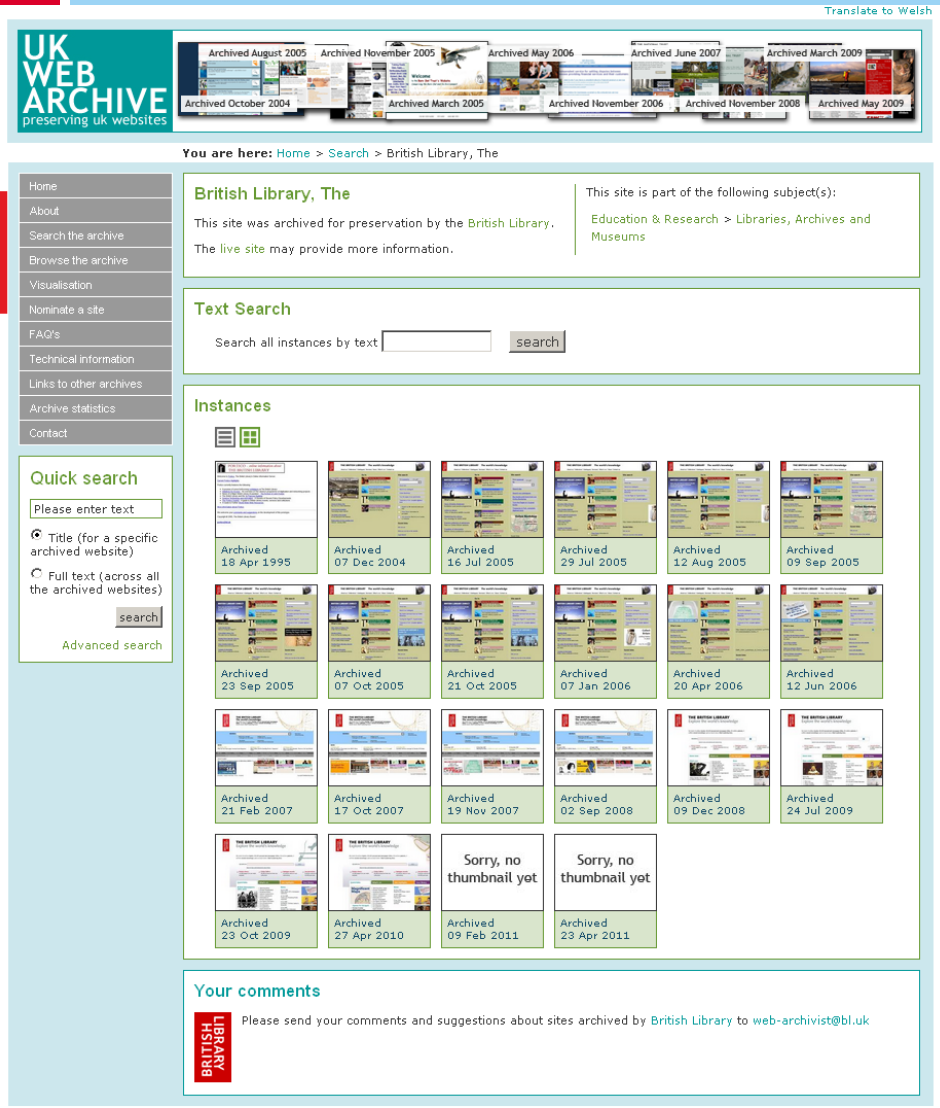
- Survey of web archiving initiatives (Daniel Gomes et al 2010)
  - 42 web archiving initiatives across 26 countries since 1996
  - 11 (26%) carry out broad domain crawls
  - 6.6PB of archived web resources
- How much of the Web is Archived (Scott Ainsworth et al, 2012)
  - Figures do not include web archives with restricted access ("dark archives")
  - Some parts of the web better preserved than other; some lost

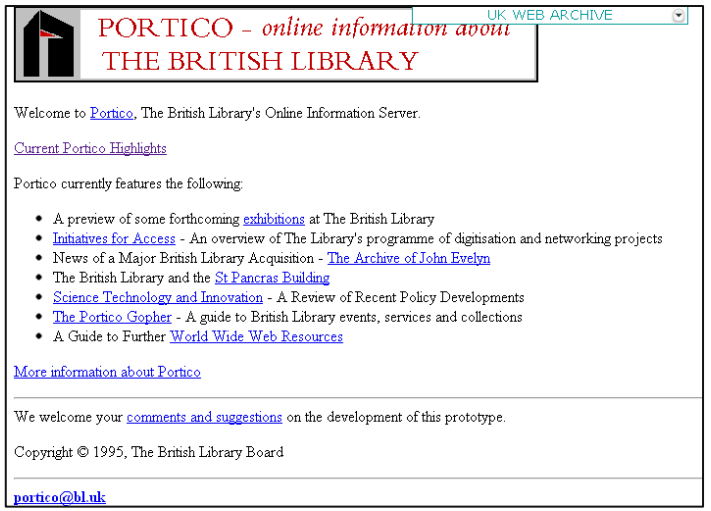
Percentage archived	# of copies in public archive
35% -90%	At least one
17-49%	2-5
1%-8%	6-10
8%-63%	>10

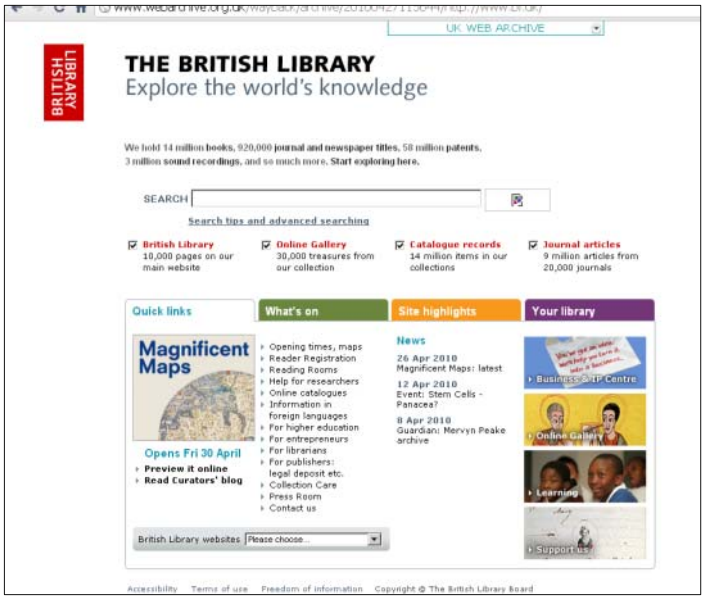
How often are web archives used?

- Little evidence of (scholarly) use of web archives
- No agreed way of calculating / benchmarking access statistics
- Archiving institutions' focus on data collection, not usage
- 19 of 29 IIPC members' archives (listed on website) have full or partial online access, often permission-based
- Large scale national web archives have restricted access – "dark" archives
  - eg Danish National Web Archive, over 280TB
    - online access for researchers with PhD or higher level
    - 20 users since 2005
- "Document-centric" access methods

Web archive as historical document







Access methods (an overview)

- IIPC members' archives has 29 entries
- URL search is the standard, universal access method - requires users to know the URL of the website they are looking for
- For many archives, full-text search is the next challenge on the roadmap

URL search	Keyword search	Full-text search	Thematic Collections	Subject Browsing	Alphabetical browsing
26	15	11	11	9	14

The UK Web Archive

- Permission-based selective archiving since 2004
  - 30% success rate
  - 14,041 websites, 58,692 instances, ~17TB WARCs
- Domain crawl from 12 April 2013 to implement non-print legal deposit
  - Expected to crawl between 4-5 million UK websites
  - Access restricted to 6 deposit libraries' reading rooms

http://www.webarchive.org.uk

Access statistic 1st April 2012 – 31 March 2013

Audience Overview

1 Apr 2012 - 31 Mar 2013

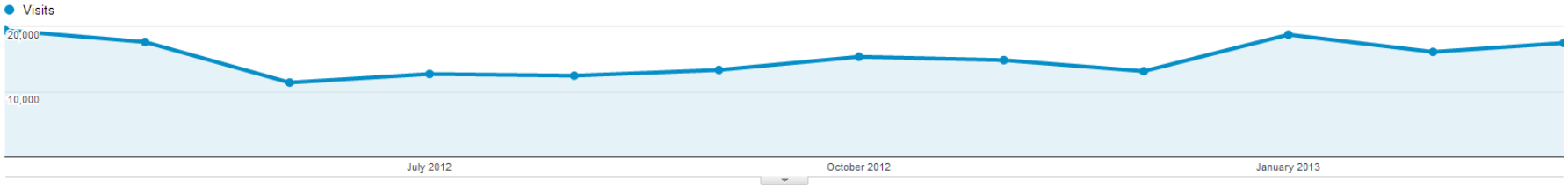
Advanced Segments | Email | Export | Add to Dashboard | Shortcut

% of visits: 100.00%

Overview

Visits vs. Select a metric

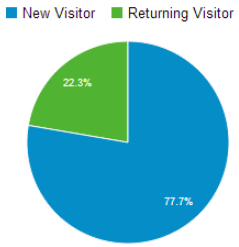
Hourly | Day | Week | Month



143,924 people visited this site

Visits	Unique Visitors	Pageviews	Pages / Visit
182,761	143,924	730,101	3.99

Avg. Visit Duration	Bounce Rate	% New Visits
00:02:20	54.79%	77.63%



Demographics
Language
Country/Territory
City
System
Browser
Operating System
Service Provider

	Language	Visits	% Visits
1.	en-us	118,980	65.10%
2.	en-gb	34,703	18.99%
3.	en	3,960	2.17%
4.	fr	2,376	1.30%
5.	ru	1,946	1.06%
6.	de-de	1,868	1.02%
7.	pl	1,472	0.81%

UK Web Archive

SCHOLARLY FEEDBACK

Scholarly feedback

- User Survey in 2012 to identify scholarly value of the UK Web Archive, as perceived by researchers
  - To obtain feedback on the access mechanisms currently offered by archive
  - To identify gaps in terms of content coverage
  - To obtain insight into reason why researchers may or may not use the web archive

Methodology

- By IRN Research between May and June 2012
- 94 telephone interviews with previous and non-users of the UK Web Archive – 74% are non-users
- A small group undertook a second phase, running search and detailing each stage – documented as case studies

Subject	Non-users	Users
Arts and Humanities	33	10
Social Sciences	27	11
Science Technology Medicine	4	3
Total	64	24
Unclassified	6	-

Scholarly value

Non users	Users
Appreciate potential value but for many no relevant content	All understand the value as snapshot of selective sites at specific times
More special collections would increase value	Value would increase with more scientific and technical content

Access Mechanisms

Non users	Users
Search tool easy to use but complicated for minority	Majority satisfied with presentation of results and ease of use of site
Most search / browse by special collections	More interest in visualisation tools
Search results unstructured and random	Need for improved data mining tools
More explanation about functions and features needed	
Limited interest in visualisation tools	

Content coverage

Non users	Users
More relevant special collections	More images, illustrations, rich media
More images, blogs	Politics, contemporary British history
Current content not relevant	Too much missed from specific websites

Additional functions and features

Non users	Users
Improvements to search results pages	6-monthly updates
Interactive features	Interactive features
Facility to suggest special collections	
More information regarding selection policy	
Too much text on home page	

Why do researcher use / not use a web archive

- Relevance of content determines whether researchers use a web archive
- Selective web archives please some but disappoint others
- Still a significant target group within the research community yet to be reached

SCHOLARLY USE OF WEB ARCHIVES

Scholarship is changing

- Blurred boundaries between scholarly sources and popular sources, even more so in the context of the web
- Any source used for scholarly purposes can be defined as scholarly source
- Scholarship is evolving: computational engaged research gaining momentum eg digital humanities
  - Redrawing disciplinary boundaries
  - Less text-based, multi-media driven
  - Web playing an important role – will archives of the web too?

Scholarly use (of digital sources): key characteristics

- Availability or accessibility
- Text and paratext, defined by Gérard Genette as “accompaniment” that “surround or prolong the text”. Niels Brugger (2010) applied this concept to websites and argues it is different in form and function, and plays a crucial role in textual coherence of a website
- Or context, in the usual sense of the word, eg out and in-links
- Citation – backbone of research - requires persistence identification of sources, ideally retrievable
- Sources relevant and specific to research question, without any arbitrarily imposed (national , geographical or format related) boundaries
- Quality – conforming to certain expectation or authoritativeness
- Flexibility /ability to apply digital methods for analytics and discovery of new knowledge

Requirements for web archives

Characteristics of Scholarly use	Requirements for web archives
Availability	No access restriction, available online
Paratext and context	Access to collection policy and scope, crawl configuration, craw log and any contextual information
Persistence and citability	- Longevity of web archives - Persistent identifiers - Standards of citing archived websites - Integration with bibliographical management tools (eg Zotero)
Collect / organise research corpus	- Archiving of research corpora on demand -Means to mix and match and reassemble corpora based on research questions -Social network content
Quality	- Archival version represents as much as possible the live website in completeness, intellectual content, behaviour and look and feel -Curation
Applying Digital methods	- Multiple access methods including data analytics and visualisations - Access to web archives as “big data”
Boundary & format-independent	- Interlinked web archives - integration with other digital and printed holdings eg books, ejournals

Unique Selling Points (USPs)

- The live web as an fast evolving, interactive, multi-dimensional, open and participatory and interlinked collective system
- Web archives as static, flat, exclusive, individual systems with boundaries and limitations
- We cannot compete with the live web (not should we); Law change and archiving technology improvement take time
- Focus on USPs – things that differentiate web archives from the live web
  - Some web resources have vanished and web archives hold the only copies of these
  - Periodic snapshots showing evolution and change of websites
  - Web archives as comprehensive historical datasets - lends itself to opportunities for analytical access
- Think out of the box (eg. archiving social network making use of APIs)

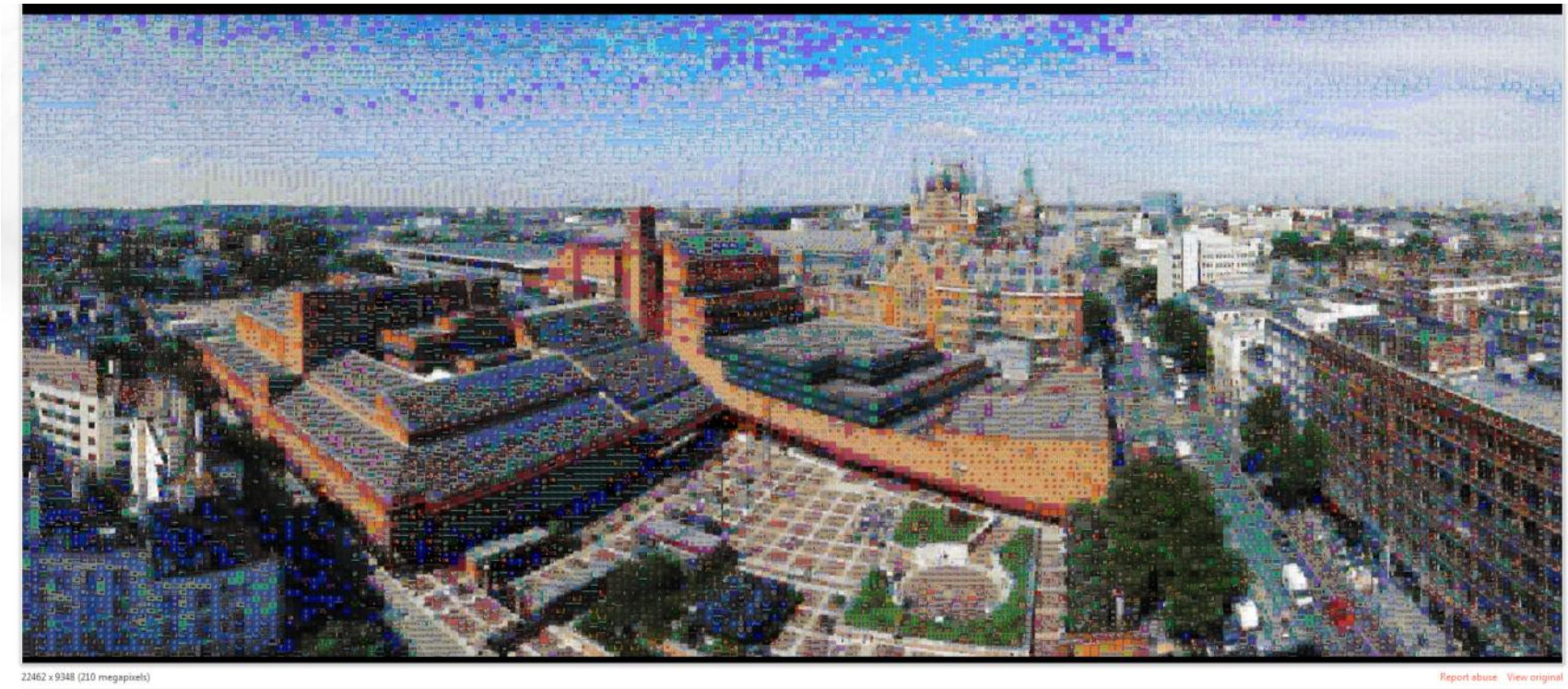
Analytical access –discovering value of the haystack

- Shift of focus from the level of single webpages or websites to the entire web archive collection or multiple archives
- Support survey, annotation, contextualisation and visualisation
- Allows discovery of patterns, trends and relationships
- The “big data” approach to analysing and using web archives
  - Added dimension: time
- Helps addresses a number of challenging issues for web archiving: scalability, components missed by crawlers
- Issues
  - Scepticism/suspicion about ‘hidden’ algorithms
  - Biases in the data
  - Managing expectation: analytical tools finished products or first steps?
  - Ethical /privacy issues

Conclusion

- The web changes; scholarship practice and methods change too
- Web archives are parts of the live web
- The web is too big for any single organisation to preserve – web archives need to join up
- Web archived can be used for references as well as analytics
- Restricted access undermines the value of web archives but there is plenty we can do to bring web archives to the scholars
  - Restriction mostly on providing access to the “text”
  - Highlight our USPs
  - Fit in with researchers’ workflow – how they do research
  - Full potential of web archives are yet to be exploited

Showing the big picture



http://seadragon.com/view/wky