

Archiving the UK Web

Digital Preservation: What I Wish I Knew Before I Started
24 January 2012

Helen Hockx-Yu

Head of Web Archiving
British Library

Web archiving: the basics

■ What

- Selecting, capturing, storing, preserving and managing access to snapshots of websites over time

■ How

- Use crawler software to download websites automatically
- Selective or domain archiving
- Provide access in a Web Archive

■ When

- Since mid 1990s

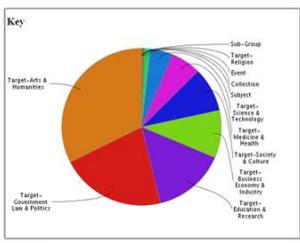
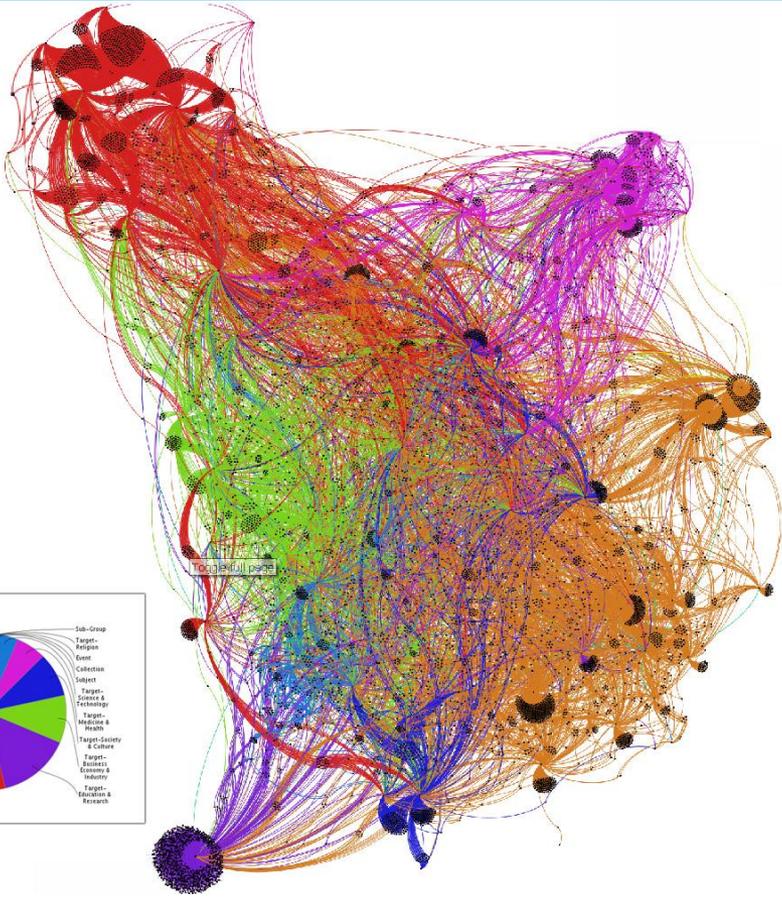
■ Who

- Heritage and memory organisations, eg (IIPC)
- University libraries
- Not-for-profit and commercial organisations, eg Internet Archive
- Individual researchers

■ Why

- Global information resource
- Artefact of cultural and technology change
- Representative sample of the web: historical and sociological data that may not be found elsewhere
- Part of national digital heritage - legal requirements

Scale: needle and haystack



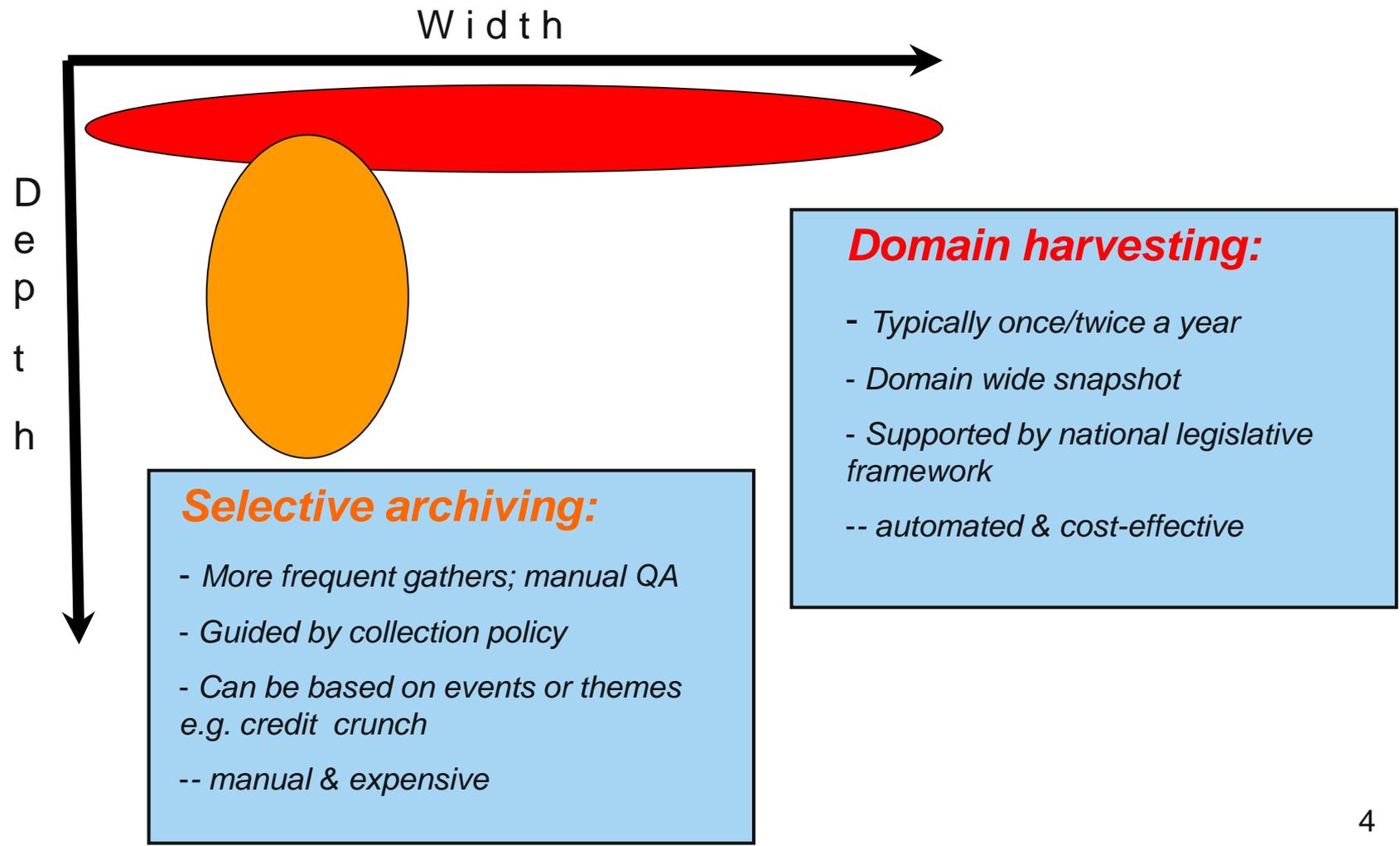
- Google: “seen 1 trillion unique URLs”
- more than a billion new pages are added to the web every day
- The UK web domain
 - 9 million .uk domain names registered in December 2010
 - ~ 1 million using other domain names
 - Growing at 11% - 14% per year
 - 40% estimated to be in scope for Legal Deposit
 - Estimated ~110TB each UK domain crawl

Subject hierarchy visualisation [UK Web Archive](#)

- over~ 10,000 websites collected since 2004
- ~ 41,000 instances

Selective versus domain archiving

- Two complementary approaches: selective and domain archiving



Key processes of web archiving

- Selection – decide what websites to archive and to include as part of a web archive collection
 - [The British Library Collection Development Policy for Websites](#)
- Automated downloading of selected websites using crawler software
- Storage
 - Archival format, eg WARC
- Access and use
 - Currently 3 ways to access the UK Web Archive
 - <http://www.webarchive.org.uk/>
 - Catalogue records (at special collection level)
 - Integrated search (Primo)
- Digital preservation
 - Common and hard problem faced by all – work with the experts
 - Bit-level preservation – long term integrity of ingested bits
 - Describe the digital objects we have: metadata profile, WARC, document original (technical) environment of websites

Web archiving paradoxes

- Small, closed community of practitioners – need research & reaching out to other communities
- Doubts and scepticism from various quarters
- Traditional “document-centric” approach does not scale up - canonical mission of heritage institutions being challenged
- Many technical challenges – the constant need to respond to the evolving web
 - Harvests are at best snapshots or samples
 - cannot get everything: resource and legal constraints; robot.txt exclusion, protected content
 - do not get every version: rate of change
 - the issue of temporal consistency
 - Crawler works well with HTML but struggles to capture advanced web content, e.g. rich media, dynamic and interactive content
 - “Bad” content
 - search engine spam, scam / malwaresites
 - Inadvertent ‘traps’
 - Illegal content
 - Rendering software does not always “replay” the archived content
 - Cannot replay streaming media
 - “live leakage”
- Access problem
 - Restricted access
 - Where are the users and what do they want?
- Legal issues
 - Risks of “republishing” – libel, copyright
 - Legal Deposit offers some protection but access restricted to premises of LD institutions

Web archive as historical documents

Translate to Welsh



You are here: Home > Search > British Library, The

- Home
- About
- Search the archive
- Browse the archive
- Visualisation
- Nominate a site
- FAQ's
- Technical information
- Links to other archives
- Archive statistics
- Contact

Quick search

Please enter text

Title (for a specific archived website)

Full text (across all the archived websites)

Advanced search

British Library, The

This site was archived for preservation by the British Library.
The live site may provide more information.

This site is part of the following subject(s):

Education & Research > Libraries, Archives and Museums

Text Search

Search all instances by text

Instances

Archived 18 Apr 1995	Archived 07 Dec 2004	Archived 16 Jul 2005	Archived 29 Jul 2005	Archived 12 Aug 2005	Archived 09 Sep 2005
Archived 23 Sep 2005	Archived 07 Oct 2005	Archived 21 Oct 2005	Archived 07 Jan 2006	Archived 20 Apr 2006	Archived 12 Jun 2006
Archived 21 Feb 2007	Archived 17 Oct 2007	Archived 19 Nov 2007	Archived 02 Sep 2008	Archived 09 Dec 2008	Archived 24 Jul 2009
		Sorry, no thumbnail yet	Sorry, no thumbnail yet		
Archived 23 Oct 2009	Archived 27 Apr 2010	Archived 09 Feb 2011	Archived 23 Apr 2011		

Your comments

Please send your comments and suggestions about sites archived by British Library to web-archivist@bl.uk

UK WEB ARCHIVE

PORTICO - online information about THE BRITISH LIBRARY

Welcome to [Portico](#), The British Library's Online Information Server.

[Current Portico Highlights](#)

Portico currently features the following:

- A preview of some forthcoming [exhibitions](#) at The British Library
- [Initiatives for Access](#) - An overview of The Library's programme of digitisation and networking projects
- News of a Major British Library Acquisition - [The Archive of John Evelyn](#)
- The British Library and the [St Pancras Building](#)
- [Science Technology and Innovation](#) - A Review of Recent Policy Developments
- [The Portico Gopher](#) - A guide to British Library events, services and collections
- A Guide to Further [World Wide Web Resources](#)

[More information about Portico](#)

We welcome your [comments and suggestions](#) on the development of this prototype.

Copyright © 1995, The British Library Board

portico@bl.uk

UK WEB ARCHIVE

THE BRITISH LIBRARY

Explore the world's knowledge

We hold 14 million books, 920,000 journal and newspaper titles, 58 million patents, 3 million sound recordings, and so much more. Start exploring here.

SEARCH

[Search tips and advanced searching](#)

<input checked="" type="checkbox"/> British Library 10,000 pages on our main website	<input checked="" type="checkbox"/> Online Gallery 30,000 treasures from our collection	<input checked="" type="checkbox"/> Catalogue records 14 million items in our collections	<input checked="" type="checkbox"/> Journal articles 9 million articles from 20,000 journals
--	---	---	--

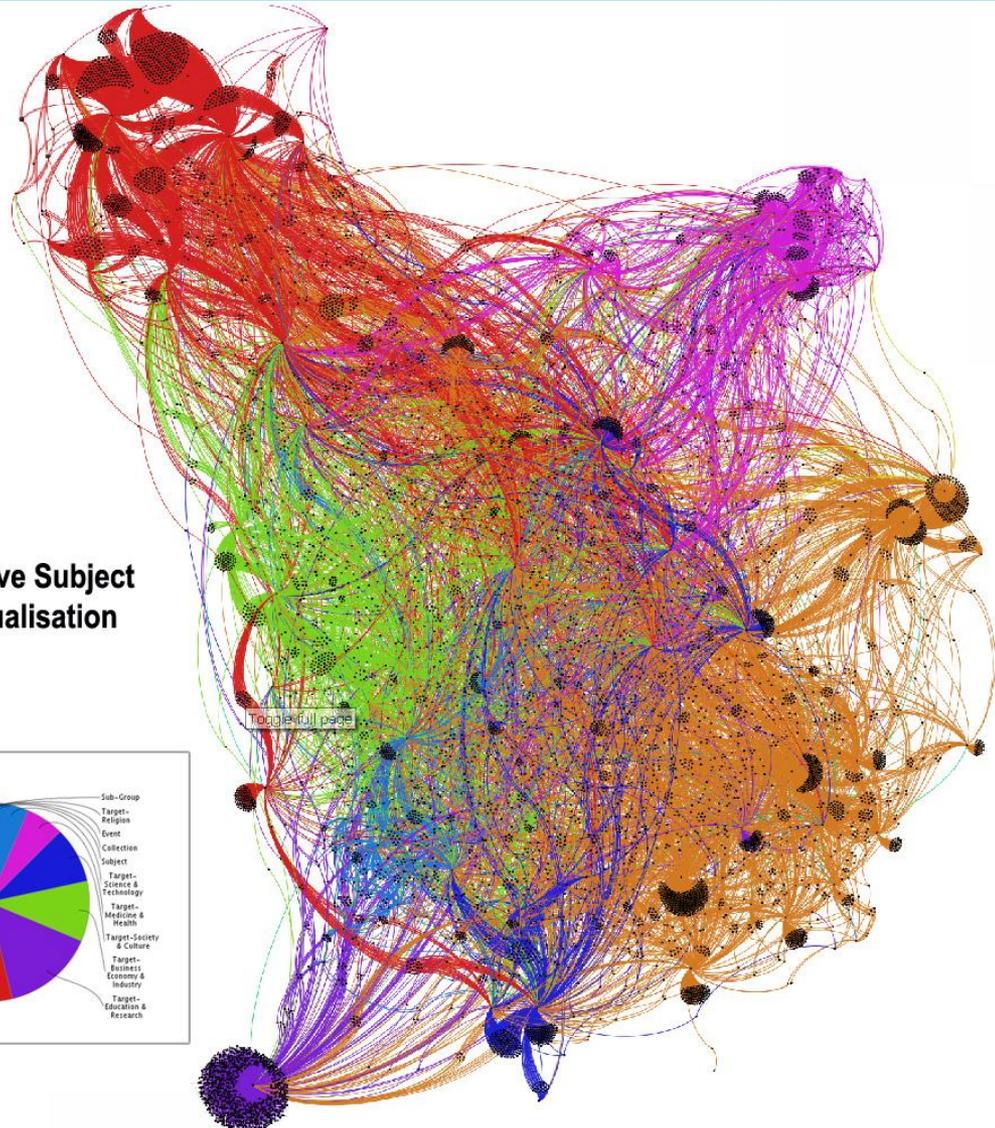
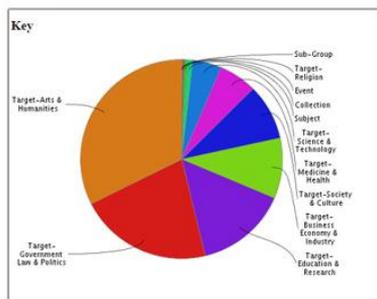
Quick links	What's on	Site highlights	Your library
<p>Magnificent Maps</p> <p>Opens Fri 30 April</p> <p>Preview it online</p> <p>Read Curators' blog</p>	<p>Opening times, maps</p> <p>Reader Registration</p> <p>Reading Rooms</p> <p>Help for researchers</p> <p>Online catalogues</p> <p>Information in foreign languages</p> <p>For higher education</p> <p>For entrepreneurs</p> <p>For librarians</p> <p>For publishers: legal deposit etc.</p> <p>Collection Care</p> <p>Press Room</p> <p>Contact us</p>	<p>News</p> <p>26 Apr 2010 Magnificent Maps: latest</p> <p>12 Apr 2010 Event: Stem Cells - Panacea?</p> <p>8 Apr 2010 Guardian: Mervyn Peake archive</p>	<p>Business QIP Centre</p> <p>Learning</p> <p>Support us</p>

British Library websites

Accessibility | Terms of use | Freedom of information | Copyright © The British Library Board

The value of the haystacks – content visualisation

UK Web Archive Subject Hierarchy Visualisation



The value of the haystacks - N-Gram search

[Translate to Welsh](#)



You are here: [Home](#) > [UK Web Archive N-gram](#)

Provided by:



- Home
- About
- Search the archive
- Browse the archive
- Visualisation
- Nominate a site
- FAQ's
- Technical information
- Links to other archives
- Archive statistics
- Contact

Quick search

Please enter text

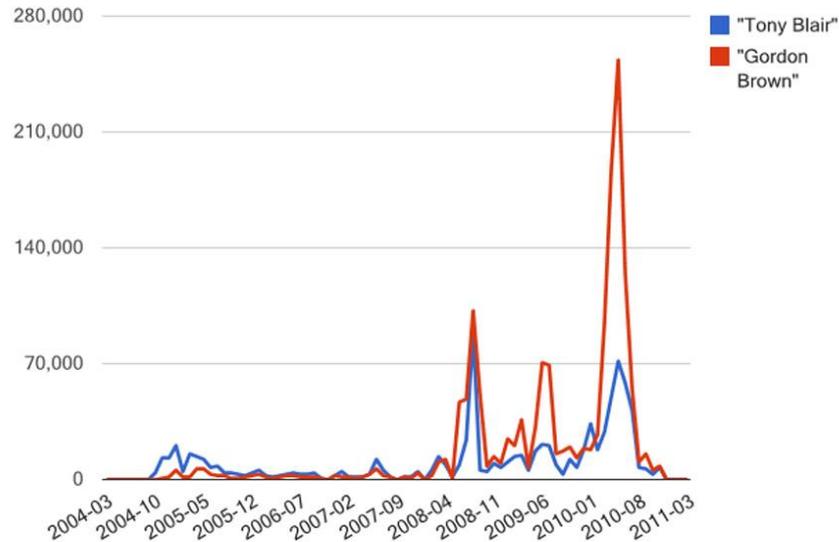
- Title (for a specific archived website)
- Full text (across all the archived websites)

search

[Advanced search](#)

UK Web Archive N-gram

UK Web Archive N-gram visualisation



Conclusion

- 14 years of web archiving – significant progress
- Yet plenty of scope for further development
- Look beyond current practices and take advantage of technologies designed for live web
- Shift of focus
 - From single page or site to entirety of web archive collection - not just for reference but also for analytics
 - Human to machine access
- Continue to grow the UK Web Archive
 - Representative of UK domain
 - used for scholarly research in a range of disciplines
 - known as the place where researchers and general public look for inactive and/or historical versions of UK websites