

# Scholarly Use of Web Archives

Helen Hockx-Yu  
Head of Web Archiving  
British Library

15 February 2013

# Web Archiving initiatives worldwide



## 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)
  - Regards search engines as one category of 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?

- 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
- No agreed way of calculating / benchmarking access statistics
- Little evidence of scholarly use of web archives, making it difficult to understand requirements

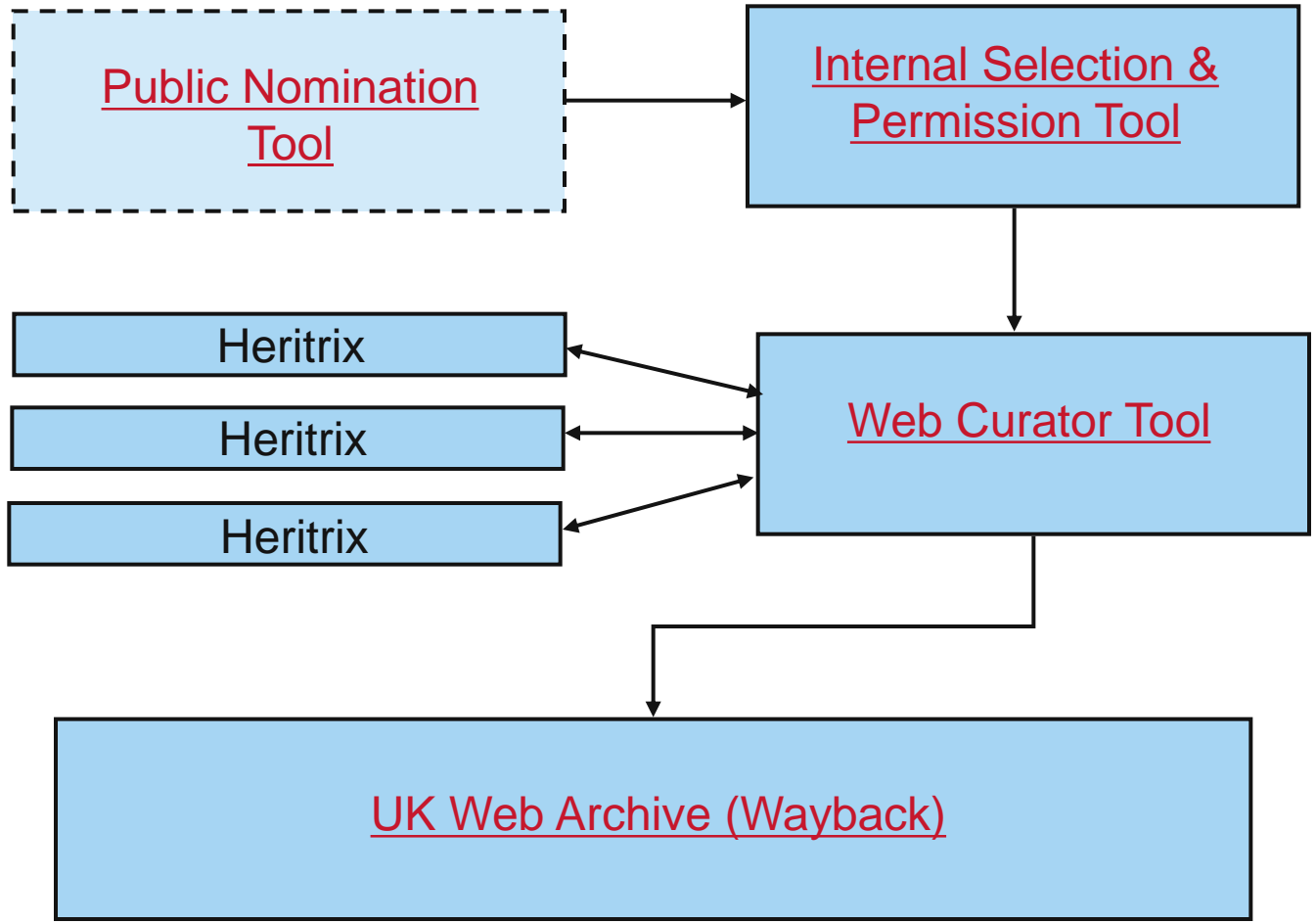
# The UK Web Archive

- Websites archived by British Library and partners since 2004
- Permission-based selective archive with **30%** success rate
- 131,164 websites, 54,604 instances, ~14TB WARC

The screenshot shows the UK Web Archive homepage. At the top, there's a navigation bar with the UK Web Archive logo and a series of thumbnails for archived websites with dates. Below this is a 'You are here: Home' breadcrumb. The main content area is divided into several sections: a 'Welcome to the UK Web Archive' section with introductory text, a 'Quick search' section with a search form, and an 'Explore the Special Collections' section with a grid of collection thumbnails. On the right side, there are two vertical panels: 'Quick website links' with frequently asked questions and 'Browse by Subject' with a list of subject categories. At the bottom, there's a 'New! N-gram Search' section with a small chart and a 'Notice and takedown | Terms and conditions | Privacy statement' footer.

<http://www.webarchive.org.uk>

# Selective Web Archiving Workflow



▪ *Courtesy of Dr Andrew Jackson, British Library*

# Web archive as historical document

Translate to Welsh

**UK WEB ARCHIVE** preserving uk websites

Archived August 2005, Archived November 2005, Archived May 2006, Archived June 2007, Archived March 2009, Archived October 2004, Archived March 2005, Archived November 2006, Archived November 2008, Archived May 2009

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

**British Library, The**

This site is part of the following subject(s):  
Education & Research > Libraries, Archives and Museums

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

**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		

**Quick search**

Please enter text

Title (for a specific archived website)

Full text (across all the archived websites)

Advanced search

**Your comments**

Please send your comments and suggestions about sites archived by British Library to [web-archivist@bl.uk](mailto:web-archivist@bl.uk)

**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](mailto:portico@bl.uk)

**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

- British Library**  
10,000 pages on our main website
- Online Gallery**  
30,000 treasures from our collection
- Catalogue records**  
14 million items in our collections
- Journal articles**  
9 million articles from 20,000 journals

**Quick links** | **What's on** | **Site highlights** | **Your library**

**Magnificent Maps**

- Opening times, maps
- Reader Registration
- Reading Rooms
- Help for researchers
- Online catalogues
- Information in foreign languages
- For higher education
- For entrepreneurs
- For librarians
- For publishers: legal deposit etc.
- Collection Care
- Press Room
- Contact us

**News**


- 26 Apr 2010 Magnificent Maps: latest
- 12 Apr 2010 Event: Stem Cells - Panacea?
- 8 Apr 2010 Guardian: Mervyn Peake archive

**Support**

British Library websites

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

# UK Web Archive: search interface


Provided by: 

You are here: [Home](#) > [Search](#)

- Home
- About
- Search the archive**
- Browse the archive
- Visualisation
- Mementos
- Nominate a site
- FAQ's
- Technical information
- Archive statistics
- UK Web Archive Blog
- Contact

## Advanced Search

### Full Text Search



Restrict by date From:   
 Format: yyyy-mm-dd To:

Restrict search to a Subject

Restrict search to a Special Collection

Restrict by Archiving Organisation

Group Results by Domain  None

### Title Search

Website Title  Website URL  Website Title and URL

Restrict search to a Subject

Restrict search to a Special Collection

### Search tips

Tips with ► icons open and close on click.

- Search by full text
- Search by title and / or URL
- Refine search to Subjects
- Refine search to Special Collections
- If you don't find what you are looking for



# UK Web Archive: browse interface

You are here: [Home](#) > [Browse by Collection](#)

Provided by:  
LIBRARY  
HSILIRB

- Home
- About
- Search the archive
- Browse the archive
- Visualisation
- Mementos
- Nominate a site
- FAQ's
- Technical information
- Archive statistics
- UK Web Archive Blog
- Contact

**Quick search**

Please enter text

Full text (across all the archived websites)

Title (for a specific archived website)

**Advanced search**

- Browse by Subject
- Browse by Special Collection
- Browse by Website Title

## Browse by Special Collection

Each Special Collection is a group of websites, usually more than fifty and less than four hundred, brought together on a particular theme, either 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). These have been especially compiled by curators and other subject specialist to make useful and interesting Special Collections.

Remember you can use the Search tips



19th Century English...	Blogs	British Countryside	Cornwall	Credit Crunch
Darwin 200	Energy	E-publishing Trends	European Parliament ...	Free Church
Governing the Police...	Indian Ocean Tsunami...	Latin America UK	Live Art	London Mayoral Elect...

**Browse search tips**

Tips with ► icons open and close on click.

- **What are Titles?**
- **What are Subjects?**
- **What are Special Collections?**

**Visualisation**

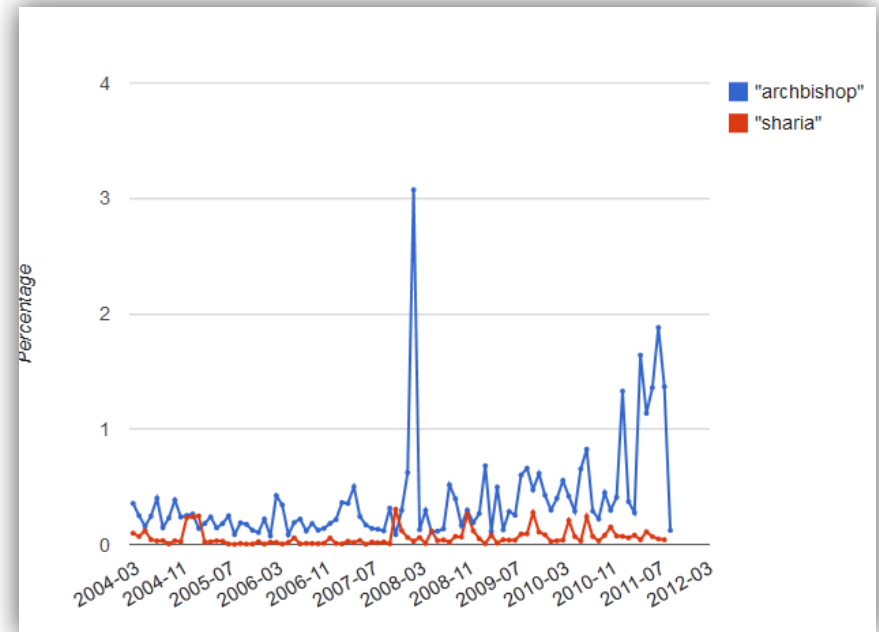
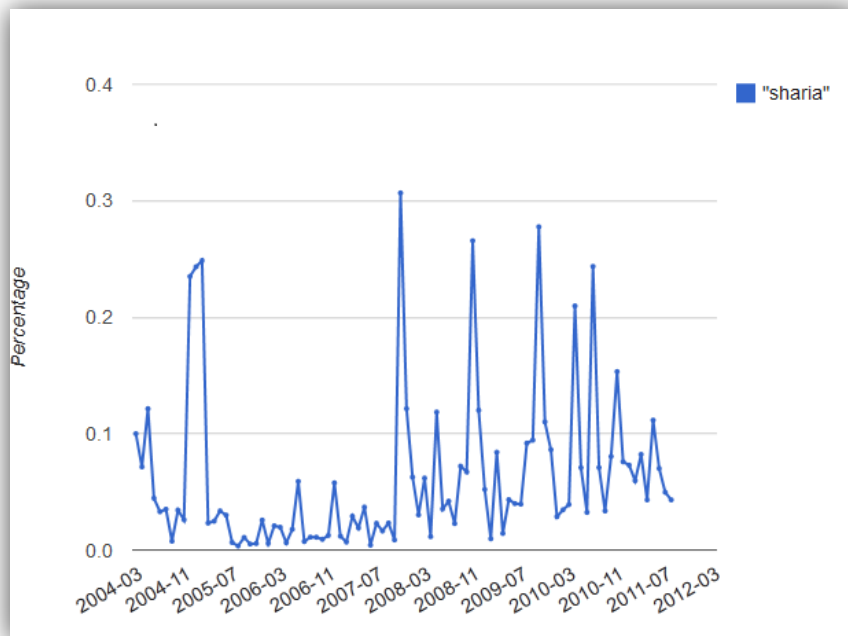
3D Wall

## Access methods (an overview)

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

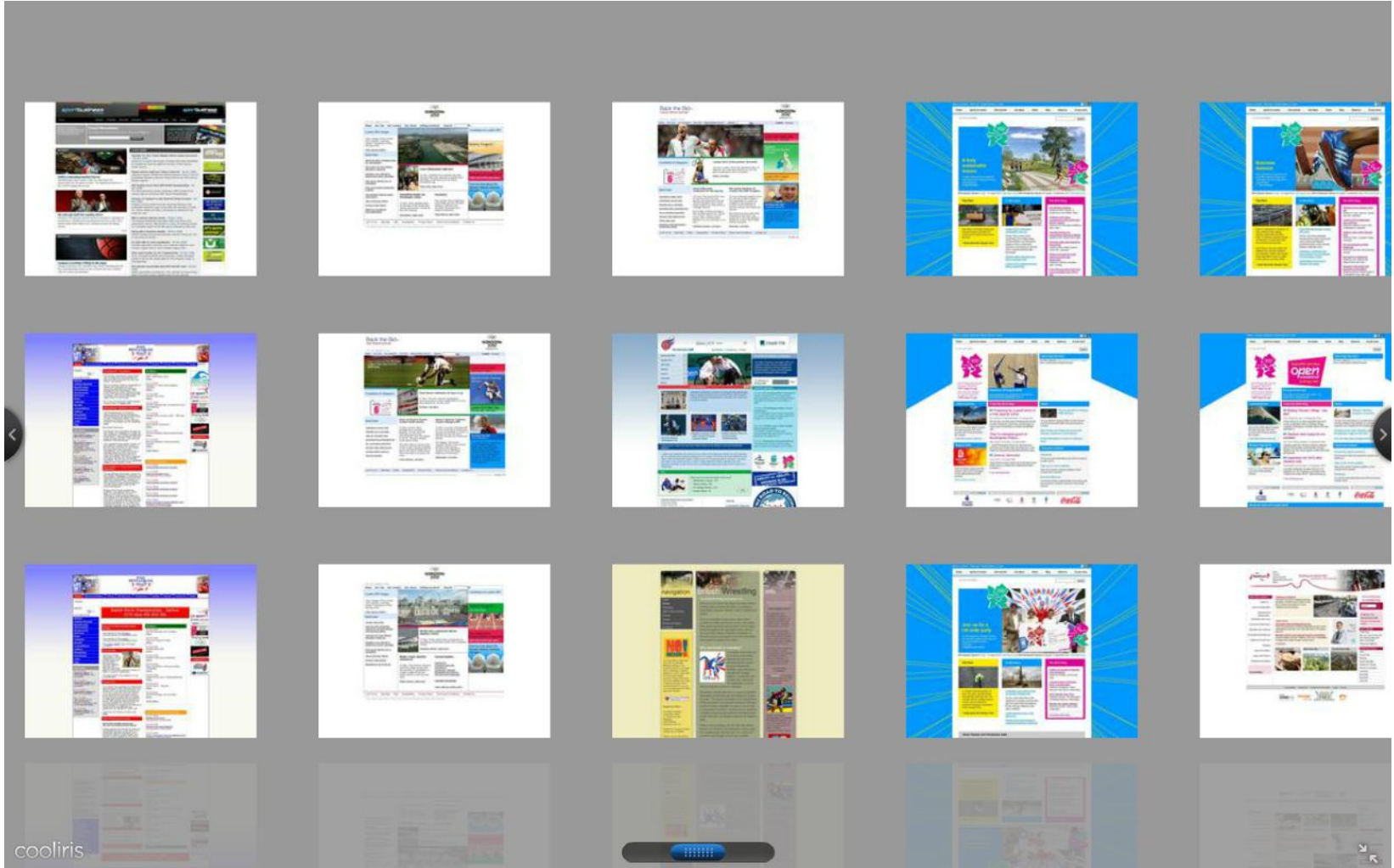
URL search	Keyword search	Full-text search	Thematic Collections	Subject Browsing	Alphabetical browsing
26	15	11	11	9	14

# Using N-gram for scholarly research



■ *Courtesy of Dr Peter Webster, Institute of Historical Research, University of London*

# UK Web Archive: visual browsing



# RSS feed of latest instances

## UK Web Archive Latest Instances

**You are viewing a feed that contains frequently updated content.** When you subscribe to a feed, it is added to the Common Feed List. Updated information from the feed is automatically downloaded to your computer and can be viewed in Internet Explorer and other programs. [Learn more about feeds.](#)

 [Subscribe to this feed](#)

## CCM - Christ Church Manchester

---

22 February 2013, 09:02:29 

## Notts and Derby Quakers

---

22 February 2013, 03:02:30 

## Transition Bro Gwaun a Community Initiative

---

22 February 2013, 01:02:48 

## Catholic Church in England and Wales

---

21 February 2013, 10:02:52 

## Luddite Link, The

---

21 February 2013, 10:02:17 

# Replacing original search function on site

by:



- Home
- About
- Search the archive
- Browse the archive
- Visualisation
- Mementos
- Nominate a site
- FAQ's
- Technical information
- Archive statistics
- UK Web Archive Blog
- Contact

## Quick search

Please enter text

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

search

**Advanced search**

## One & Other

This site was archived for preservation by the [British Library](#).  
 The [live site](#) may provide more information.  
 The Wellcome Trust commissioned a series of interviews with the participants who took part in the One & Other project. For further information please see the catalogue record on the [Wellcome Library's website](#).

This site is part of the following Special Collection(s):  
[Oral History in the UK\\*](#)  
 This site is part of the following subject(s):  
[Arts & Humanities](#)

## Text Search

Search all instances by text

## Instances

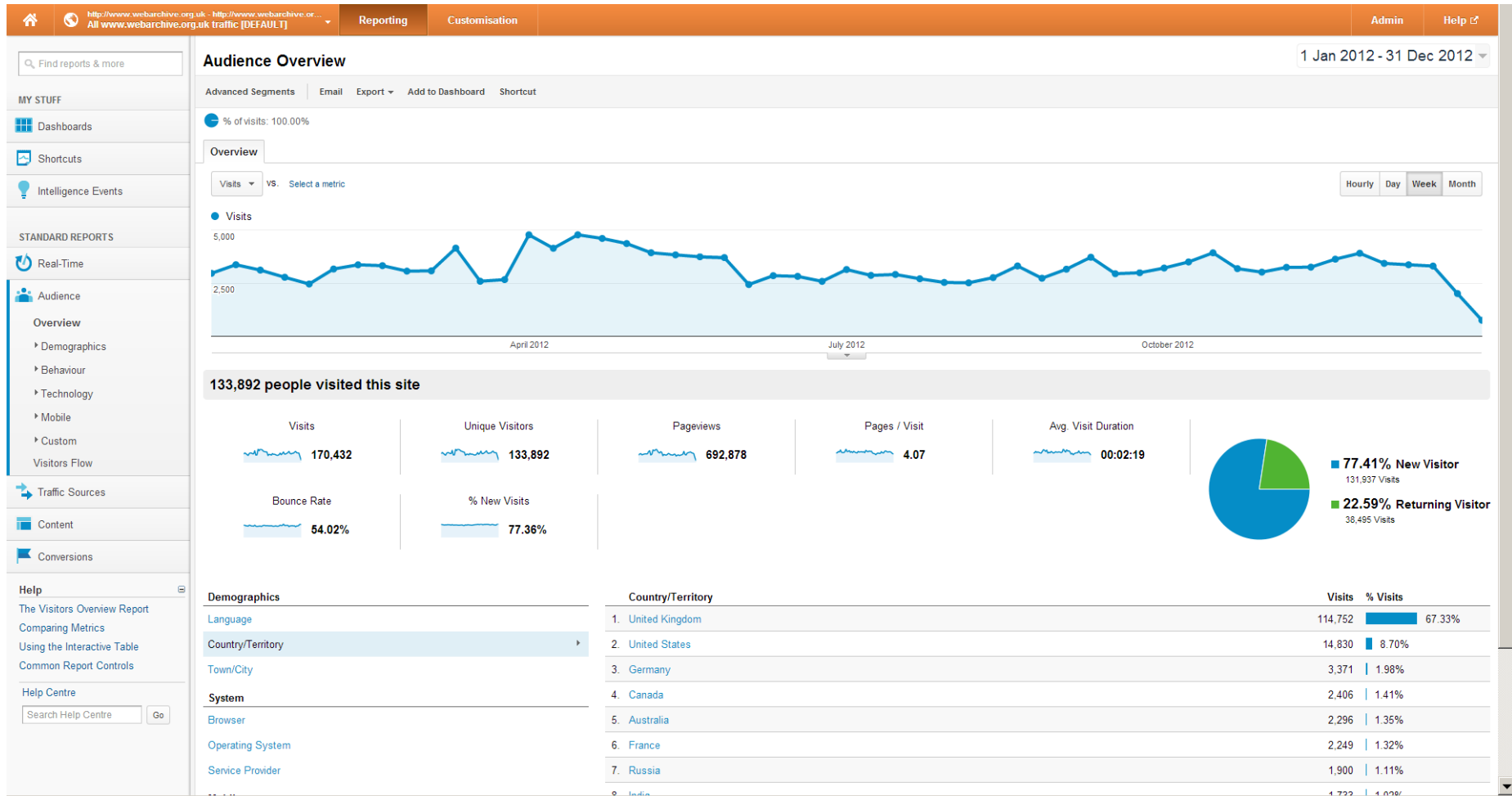


You are viewing an archival version of the website. All message boards and discussion forums have been disabled. External links, forms and search boxes do not function.



Archived  
23 Feb 2010

# Access statistic 1<sup>st</sup> Jan 2012 – 31 December 2012



# Access statistic 1<sup>st</sup> Jan 2012 – 31 December 2012

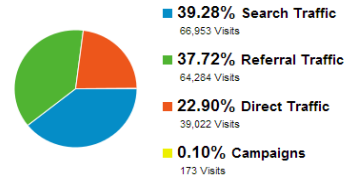
<b>Visits</b> <b>170,432</b> <small>% of Total: 100.00% (170,432)</small>	<b>Pages / Visit</b> <b>4.07</b> <small>Site Avg: 4.07 (0.00%)</small>	<b>Avg. Visit Duration</b> <b>00:02:19</b> <small>Site Avg: 00:02:19 (0.00%)</small>	<b>% New Visits</b> <b>77.36%</b> <small>Site Avg: 77.36% (0.00%)</small>	<b>Bounce Rate</b> <b>54.02%</b> <small>Site Avg: 54.02% (0.00%)</small>
---	--	--	---	--

Primary Dimension: **Source/Medium** Source Medium Other ▾

Plot Rows: Secondary dimension ▾ Sort Type: Default ▾

Source/Medium	Visits	Pages / Visit	Avg. Visit Duration	% New Visits	Bounce Rate
1. google / organic	62,386	4.04	00:02:01	78.85%	53.14%
2. (direct) / (none)	39,022	4.10	00:02:39	72.88%	56.20%
3. bl.uk / referral	5,300	5.61	00:02:52	79.49%	48.45%
4. en.wikipedia.org / referral	4,734	3.21	00:01:54	89.75%	57.67%
5. walkinginshropshire.co.uk / referral	3,079	2.07	00:02:57	79.90%	54.66%
6. northwestlondon.nhs.uk / referral	2,790	4.16	00:02:22	79.57%	40.72%
7. t.co / referral	2,301	2.79	00:01:49	67.71%	71.27%
8. ladygodiva.wordpress.com / referral	1,762	1.82	00:02:01	80.48%	72.64%
9. bing / organic	1,724	5.40	00:02:17	67.92%	40.78%
10. naturistholiday.info / referral	1,591	2.80	00:01:37	92.77%	48.02%

**170,432 people visited this site**



**Search Traffic**

Keyword

Matched Search Query

Source

---

**Referral Traffic**

Source

---

**Direct Traffic**

Landing Page

**Keyword**

Keyword	Visits	% Visits
1. (not provided)	15,817	23.62%
2. web archive	4,045	6.04%
3. one and other	1,726	2.58%
4. uk web archive	1,601	2.39%
5. welsh language board	1,426	2.13%
6. webarchive	1,319	1.97%
7. bwrdd yr iaith	864	1.29%
8. website archive	747	1.12%
9. mudeford	705	1.05%
10. bwrdd yr iaith gymraeg	609	0.91%



# Access statistic 1<sup>st</sup> Jan 2012 – 31 December 2012

Primary Dimension: Page Page Title Other ▾

Plot Rows Secondary dimension: Landing Page ▾ Sort Type: Default ▾

advanced

<input type="checkbox"/>	Page	Landing Page	Pageviews	Unique Pageviews	Avg. Time on Page	Entrances	Bounce Rate	% Exit	Page Value
<input type="checkbox"/>	1. /ukwa/	/ukwa/	57,601	43,037	00:00:55	42,982	44.30%	44.58%	\$0.00
<input type="checkbox"/>	2. /ukwa/target/32145446	/ukwa/target/32145446	22,349	16,288	00:01:05	16,287	69.91%	63.52%	\$0.00
<input type="checkbox"/>	3. /ukwa/advancedsearch	/ukwa/	14,762	7,524	00:00:26	0	0.00%	6.92%	\$0.00
<input type="checkbox"/>	4. /ukwa/target/100770/source/search	/ukwa/target/100770/source/search	9,141	6,235	00:00:50	6,189	48.73%	45.44%	\$0.00
<input type="checkbox"/>	5. /	/	6,411	4,125	00:01:03	4,125	66.98%	64.34%	\$0.00
<input type="checkbox"/>	6. /ukwa/browse	/ukwa/	5,193	3,480	00:00:13	0	0.00%	4.64%	\$0.00
<input type="checkbox"/>	7. /webarchive/	/webarchive/	3,272	2,825	00:02:13	2,825	83.15%	79.55%	\$0.00
<input type="checkbox"/>	8. /wayback/archive/20100114165728/http://oncomarchive.com/	/wayback/archive/20100114165728/http://oncomarchive.com/	3,060	2,412	00:03:59	2,398	75.52%	71.18%	\$0.00
<input type="checkbox"/>	9. /wayback/archive/20120330000303/http://www.byig-wlb.org.uk/Pages/Hafan.aspx	/wayback/archive/20120330000303/http://www.byig-wlb.org.uk/Pages/Hafan.aspx	2,874	2,259	00:00:39	2,259	34.31%	37.79%	\$0.00
<input type="checkbox"/>	10. /wayback/archive/20100223123427/http://www.oneanother.co.uk/participants/Lady-Godiva	/wayback/archive/20100223123427/http://www.oneanother.co.uk/participants/Lady-Godiva	2,699	1,955	00:03:53	1,955	72.28%	66.21%	\$0.00
<input type="checkbox"/>	11. /ukwa/info/about	/ukwa/	2,485	1,862	00:01:27	0	0.00%	25.07%	\$0.00
<input type="checkbox"/>	12. /ukwa/info/nominate	/ukwa/info/nominate	2,457	2,061	00:02:37	2,061	69.14%	64.27%	\$0.00
<input type="checkbox"/>	13. /wayback/archive/20100223121732/http://www.oneanother.co.uk/	/ukwa/target/32145446	2,145	1,408	00:00:41	0	0.00%	22.38%	\$0.00
<input type="checkbox"/>	14. /ukwa/alpha/	/ukwa/	2,077	1,226	00:00:09	0	0.00%	2.50%	\$0.00
<input type="checkbox"/>	15. /ukwa/visualisation	/ukwa/	1,810	1,135	00:00:26	0	0.00%	9.01%	\$0.00
<input type="checkbox"/>	16. /ukwa/target/15237169/source/search	/ukwa/target/15237169/source/search	1,704	1,279	00:00:55	1,279	63.02%	59.21%	\$0.00
<input type="checkbox"/>	17. /ukwa/target/66158932/source/search	/ukwa/target/66158932/source/search	1,606	957	00:00:46	957	57.47%	48.32%	\$0.00
<input type="checkbox"/>	18. /ukwa/advancedsearch	/ukwa/advancedsearch	1,596	970	00:01:29	970	50.31%	37.91%	\$0.00
<input type="checkbox"/>	19. /ukwa/target/129913/source/alpha	/ukwa/target/129913/source/alpha	1,451	769	00:00:41	769	24.97%	19.85%	\$0.00
<input type="checkbox"/>	20. /ukwa/subject/64/page/1	/ukwa/	1,441	932	00:00:16	0	0.00%	2.57%	\$0.00
<input type="checkbox"/>	21. /wayback/archive/20120330000846/http://www.byig-wlb.org.uk/English/Pages/index.aspx	/wayback/archive/20120330000303/http://www.byig-wlb.org.uk/Pages/Hafan.aspx	1,436	1,019	00:00:44	0	0.00%	12.67%	\$0.00
<input type="checkbox"/>	22. /ukwa/info/nominate	/ukwa/	1,414	1,086	00:02:37	0	0.00%	29.14%	\$0.00
<input type="checkbox"/>	23. /ukwa/ngram/	/ukwa/ngram/	1,301	479	00:00:56	478	39.54%	29.36%	\$0.00
<input type="checkbox"/>	24. /ukwa/collection/26312782/page/1	/ukwa/collection/26312782/page/1	1,138	721	00:01:38	721	38.42%	34.45%	\$0.00
<input type="checkbox"/>	25. /wayback/archive/20110630152930/http://www.barrowct.nhs.uk/index/	/wayback/archive/20110630152930/http://www.barrowct.nhs.uk/index/							

## 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 was asked to undertake 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	

## Additional functions and features

<b>Non users</b>	<b>Users</b>
Improvements to search results pages	6-monthly updates
Interactive features	Interactive features
Facility to suggest special collections	
Too much text on home page	

## Content coverage

Non users	Users
More relevant special collections	More images, illustrations, rich media
More images, blogs	Politics, contemporary British history
	Too much missed from specific websites

## Reason for using or not using UKWA

Non users	Users
Current content not relevant	Majority “very likely” to use again as there is content of interest
More information regarding selection policy	Another 39% “quite likely”
Less than a quarter “very likely” to use again	



## Why do researcher use / not use a web archive

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

collections content facility  
features home images improvements  
increase interactive interest majority  
page potential relevant results  
search selection site special  
specific suggest text tools value  
visualisation

## 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 “paratexts” to websites as objects of study: different in form and function, and play a crucial role in the 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 - non-inferiority, conformance
- 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 or context	Access to collection policy and scope, crawl configuration, crawl log and any contextual information
Persistence and citability	<ul style="list-style-type: none"> <li>- Longevity of web archives</li> <li>- Persistent identifiers</li> <li>- Standards of citing archived websites</li> <li>- Integration with bibliographical management tools (eg Zotero)</li> </ul>
Collect / organise research corpus	<ul style="list-style-type: none"> <li>- Archiving of research corpora on demand</li> <li>- Means to mix and match and reassemble corpora based on research questions</li> </ul>
Quality	<ul style="list-style-type: none"> <li>- Archival version represents as much as possible the live website in completeness, intellectual content, behaviour and look and feel</li> <li>- Curation</li> </ul>
Applying Digital methods	<ul style="list-style-type: none"> <li>- Multiple access methods including data analytics and visualisations</li> <li>- Access to web archives as “big data”</li> </ul>
Boundary & format-independent	<ul style="list-style-type: none"> <li>- Interlinked web archives</li> <li>- Integration with other digital and printed holdings eg books, ejournals</li> </ul>

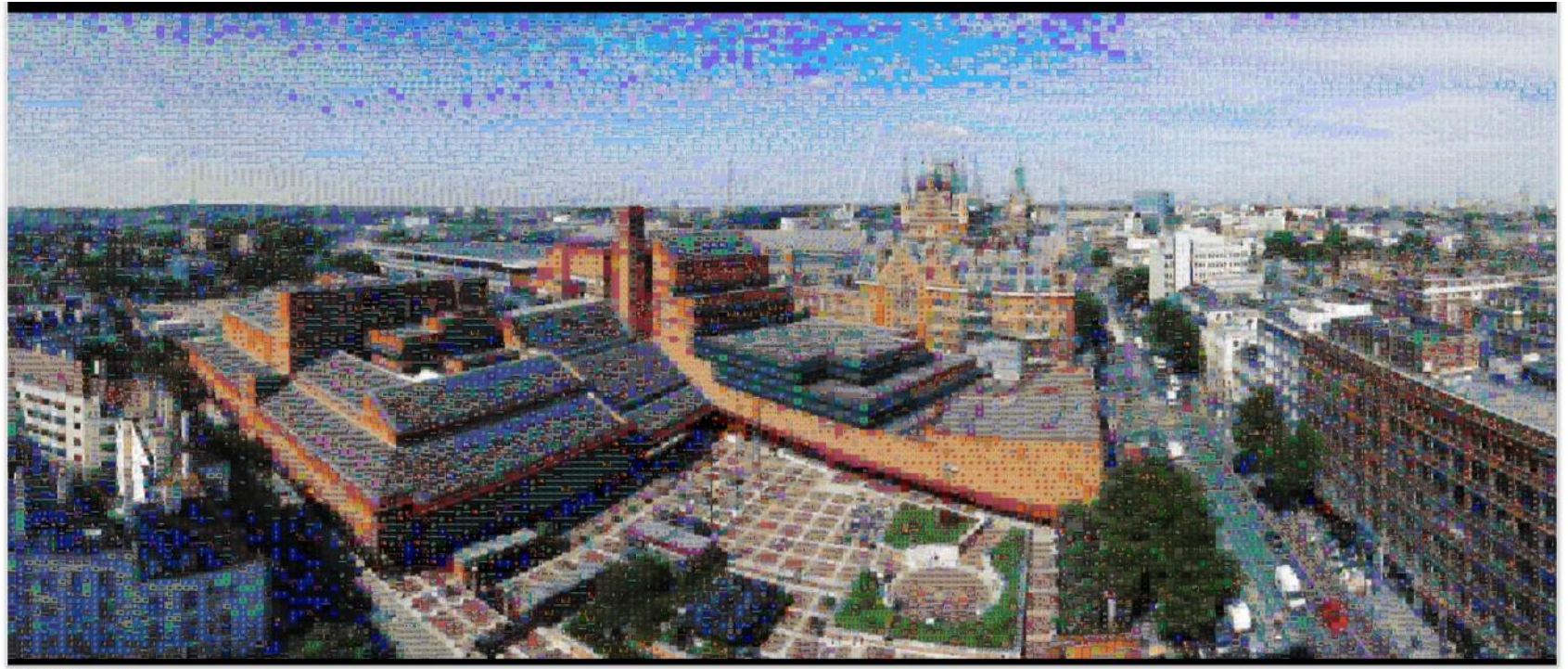
## Unique Selling Points (USPs)

- 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

# 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

# Showing the big picture



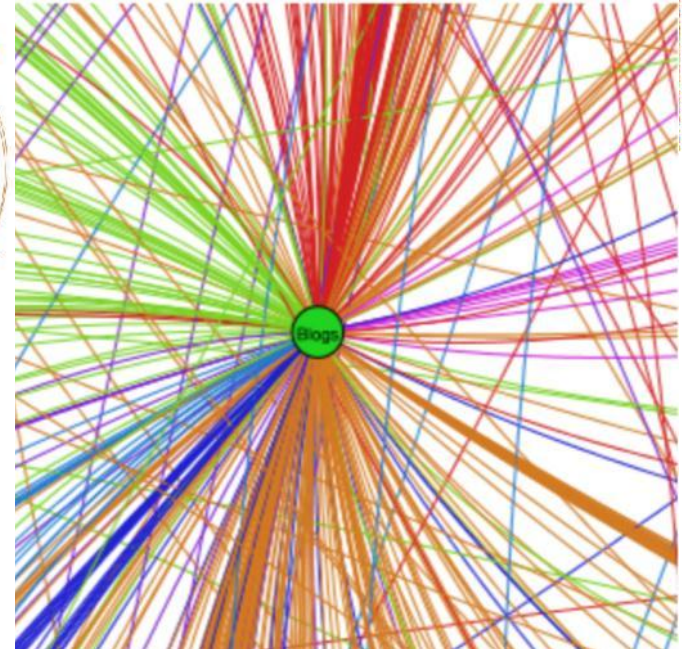
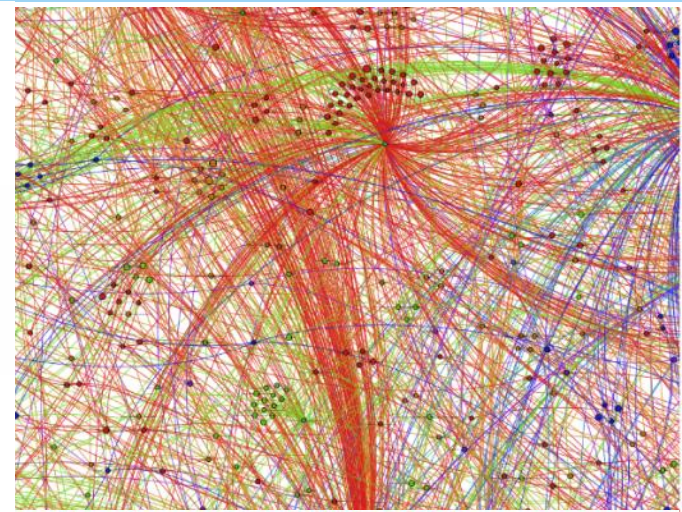
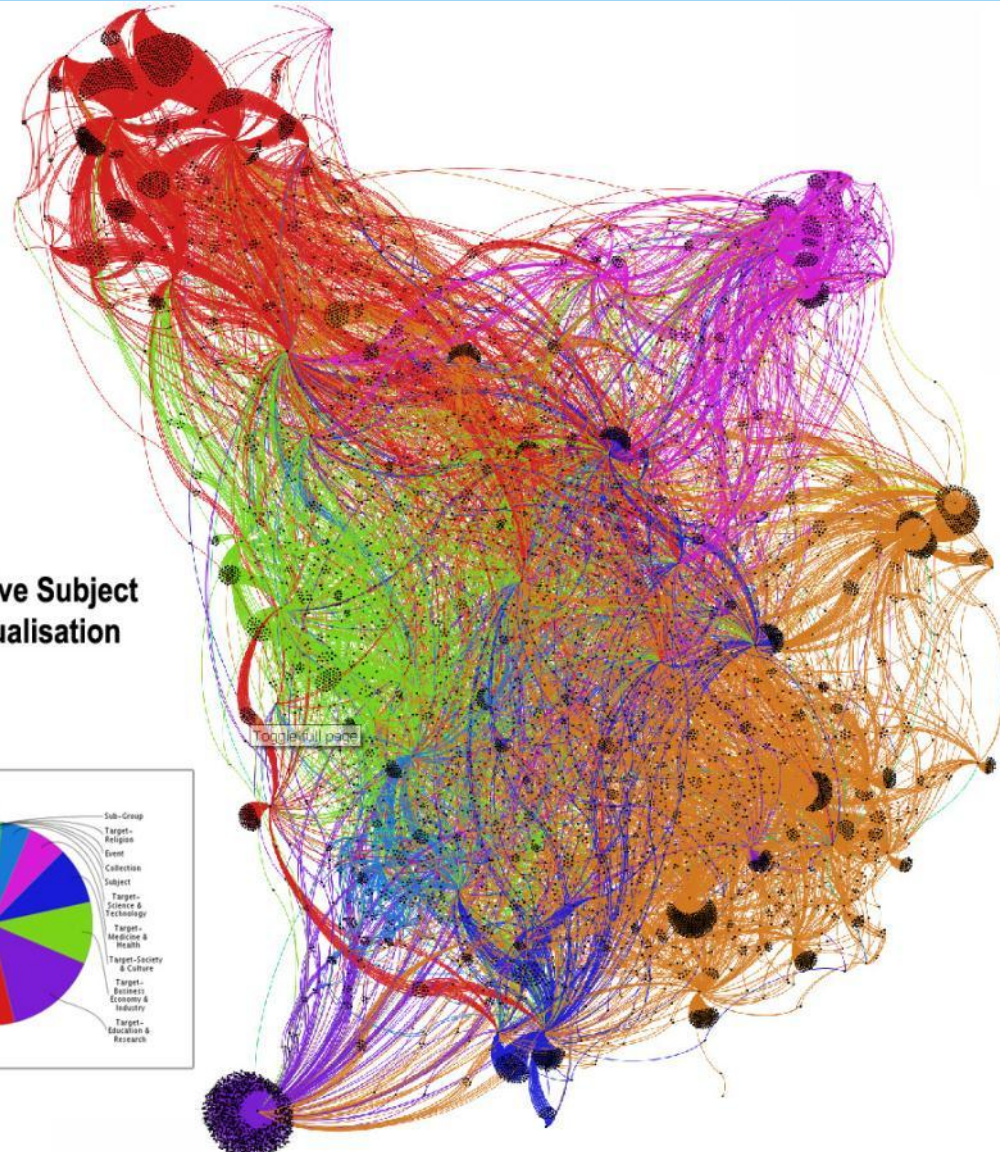
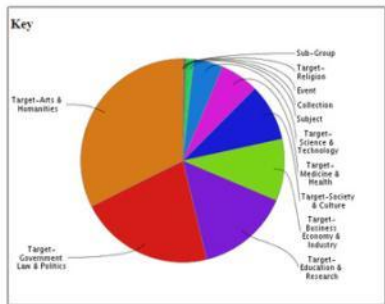
22462 x 9348 (210 megapixels)

[Report abuse](#) [View original](#)

<http://seadragon.com/view/wky>

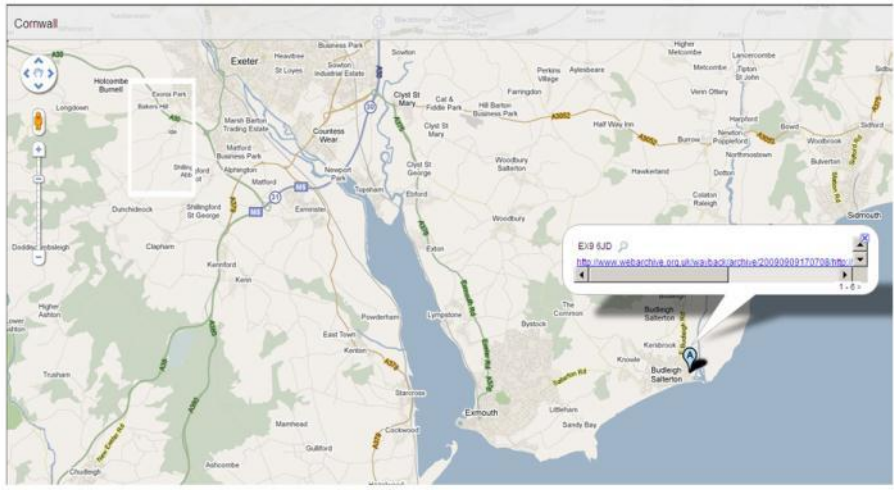
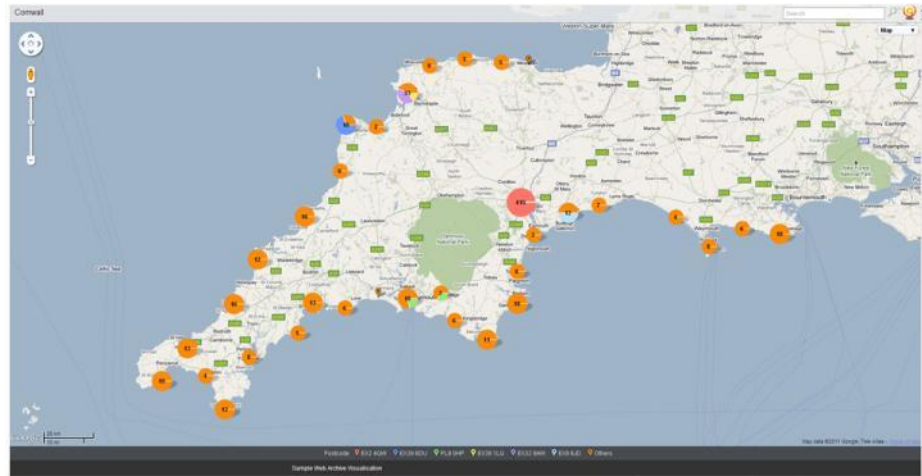
# Clustering content

UK Web Archive Subject Hierarchy Visualisation





# Postcode-based access



THE OFFICIAL GUIDE TO THE **South West Coast Path** NATIONAL TRAIL

HOME SEARCH GO THINGS TO DISCOVER WALKS FOR EVERYONE USEFUL INFORMATION

**SHORT WALKS** (up to one day)

**WALK NAME:** Salcombe to Bolt Head  
**START LOCATION:** South Sands, Salcombe  
**FINISH LOCATION:** South Sands, Salcombe  
**DISTANCE:** 3 miles (4km)  
**GRADE:** Moderate  
**TERRAIN:** Coastal footpath uneven and rocky in places; surfaced road; field footpath.  
**CIRCULAR WALK:** Yes  
**FREE FROM OBSTACLES & STEEP GRADIENTS:** No  
**RECOMMENDED BY:** [South Devon AONB Service](#)  
**WALK DESCRIPTION:**

Passing through an enchanted landscape of rocky spires and jumbled pinnacles, with inspiring views in all directions, this is a coastal walk guaranteed to lift the spirits.

The South Devon AONB Service have produced a leaflet for this walk giving directions, along with information on the wildlife, geology and history of the area. It can be viewed and printed by clicking on the 'Printer friendly page' button on the right hand side, near the bottom of the page.

To find out more about the South Devon Area of Outstanding Natural Beauty and download other walks visit their website - <http://www.southdevonaonb.org.uk>

**PUBLIC TRANSPORT INFORMATION:**  
 There are bus services to Salcombe. 606 from Kingsbridge; 92 from Plymouth and Kingsbridge. From bus stop follow the pedestrian signs to town centre and Whitestrand Pontoon.  
 For details visit [Traveline](#) or phone 0870 6082608

**NEAREST TOILETS:**  
 Public toilets at South Sands and at Whitestrand, Salcombe.

**NEAREST CAR PARKS:**  
 Shadycombe Car Park, Salcombe (Postcode for Sat Nav: **TQ8 9ND**). Tides Reach car park, South Sands.

**NEAREST REFRESHMENTS:**  
 Several pubs, cafes and restaurants in Salcombe. Tides Reach Hotel, South Sands serves refreshments and cream teas.

**FURTHER INFORMATION:**  
[Salcombe Tourist Information Centre](#) or phone 01548 843927870

**Salcombe Estuary**

**CLICK MAP TO ENLARGE**

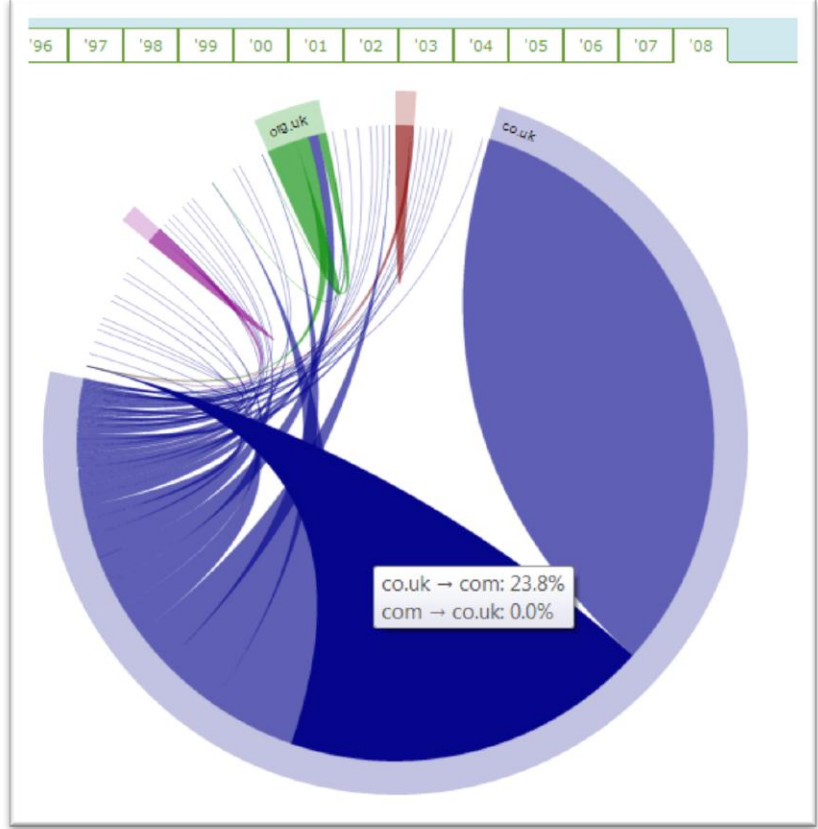
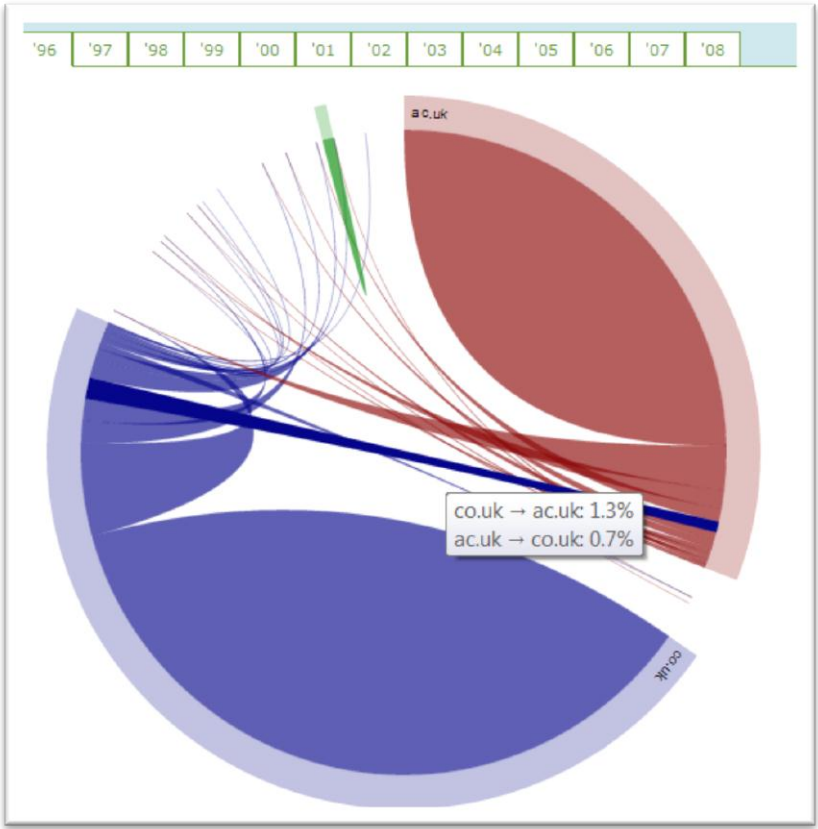
This map is reproduced from Ordnance Survey material with the permission of Ordnance Survey on behalf of the Controller of Her Majesty's Stationery Office © Crown copyright. Unauthorised reproduction infringes Crown copyright and may lead to prosecution or civil proceedings. Natural England License Number: 100049223. The Ordnance Survey mapping is included purely to provide a contextual backdrop for the walk and cannot be used for any other purpose.

# Analysing web scale data

- Internet Archive UK Domain Dataset
  - 1996-2010
  - Millions of websites
  - 2.5 billion resources
  - > 35TB

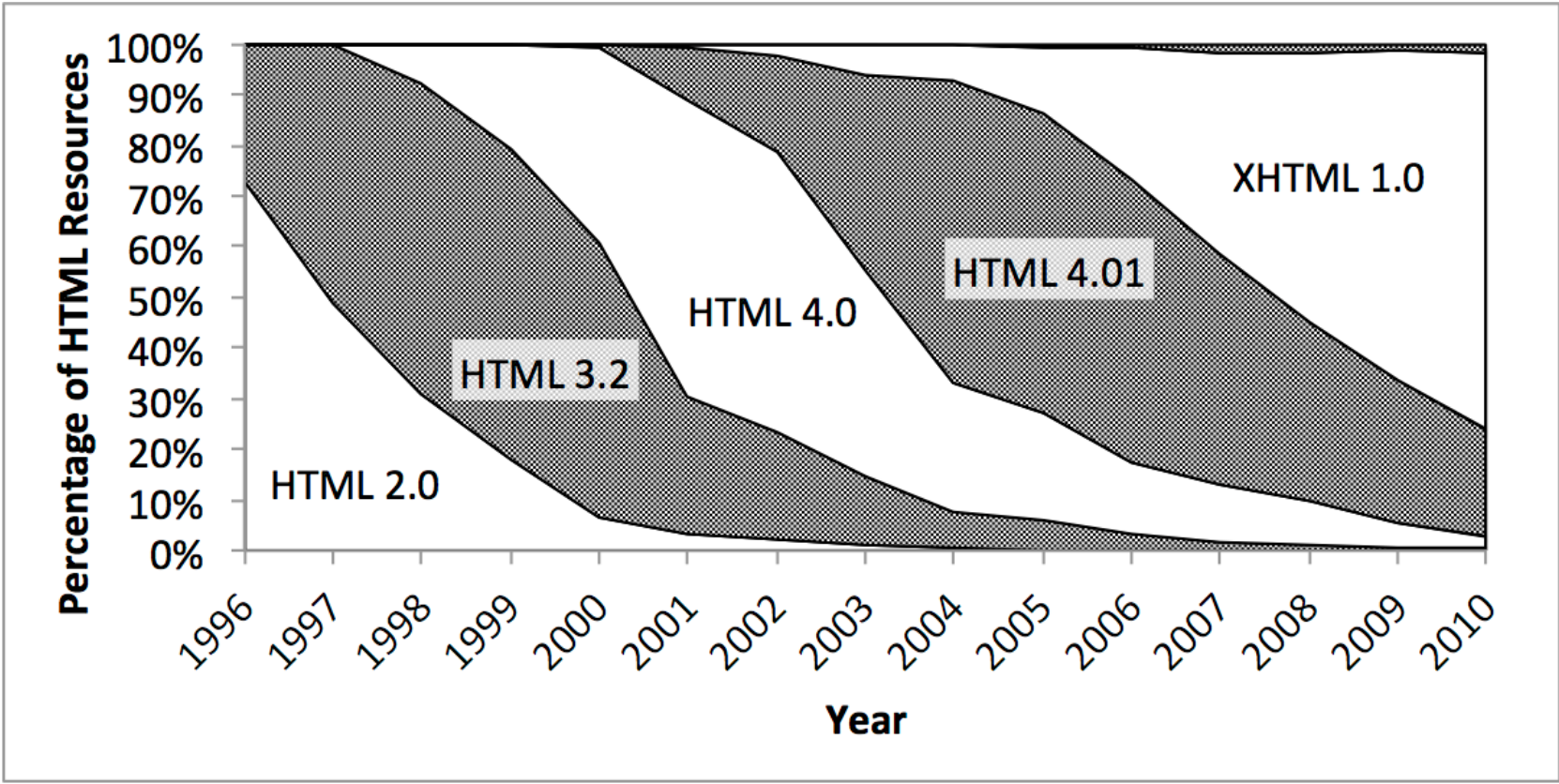


# Linkage Analysis

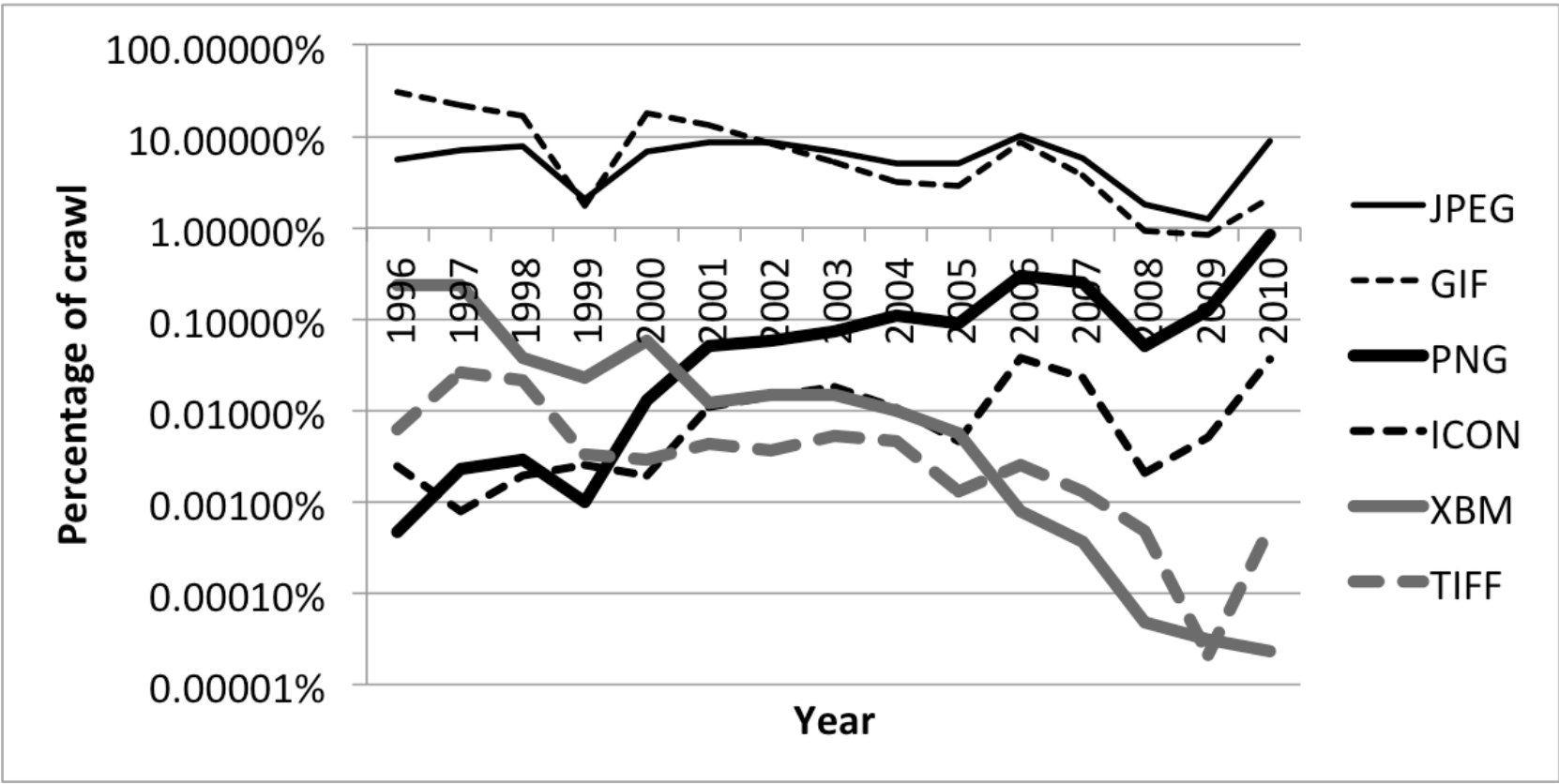


<http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/linkage>

# HTML Version Analysis

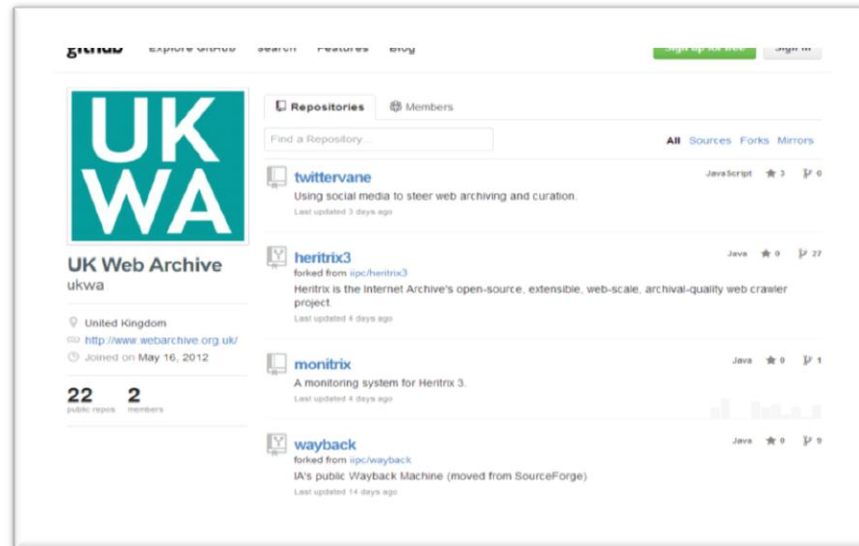


# Image Format Analysis



# Open datasets and API

- Wayback API exposing content of the UK Web Archive:  
<http://www.webarchive.org.uk/wayback/archive/xmlquery.jsp?url=http://www.bl.uk/>
- Open datasets (based on JISC UK domain dataset)
  - Geo Index
  - Format profile
  - Currently generating WAT (Web Archive Transformation) files
- Open tools



# Mementos Service

**UK WEB ARCHIVE**  
preserving UK websites

**Mementos**  
Finding historical versions of web pages (mementos) captured by multiple archives across the world.

A 'memento' is a historical version of a web resource. Memento is also a method that allows seamless discovery and delivery of individual mementos, regardless where they are held. This is based on metadata aggregated from various web archives in the world, allowing you to find the historical pages you need across several archives.

Our Mementos service exposes the Memento protocol via a simple web-based user interface, allowing you to look up which archives across the world hold mementos of any particular URL. It provides basic visualisations of this information, including breakdowns of how many mementos different archival organisations hold, and how the collection of those mementos has varied over time. Although direct browsing of the past web requires special tools or browser extensions (like Mementofox), our lookup service provides an easy way to start exploring the world wide archival history of the web.

To get started, visit the Mementos service and enter a URL, or click on one of the examples below.

**Examples**

- The history of [www.britishlibrary.org.uk](#)
- The history of [www.bl.uk](#)
- The history of [www.google.co.uk](#)
- The history of [www.bbc.co.uk](#)

**UK Web Archive** Home Mementos

http://www.langzeitarchivierung.de Find Mementos...

URL	http://www.langzeitarchivierung.de
Snapshots	60 in 1 archive(s)
Date Range	2003 to 2012 (3 months and 19 days ago)
Request Archive	via the UK Web Archive, via WebCite™

Host Chart Host Table

● archive.org

2003 2007 2012 LIVE

Snapshot Chart Snapshot Table

○ Grouped ● Stacked

● archive.org

Bookmarks: Find Mementos.

For more information, see [the Mementos homepage](#).  
This web interface uses the Memento aggregate TimeGate hosted by [lanl.gov](#).  
For more information on Memento, see [www.mementoweb.org](#).

# Improving selection - TwitterVane

Home | About

## TwitterVane

Crowd sourcing for Web Archiving

Collections Reports Streamed Tweets

**Welcome**

Selection of web archive is a manual process that relies upon people to select quality sites and nominate or submit them to web archives for inclusion. Past (national and/or international) efforts to automate selection have either failed to convince staff that automation was identifying quality resources, or have focused on providing an interface for selectors to submit sites rather than carry out the selection per se. Selection has thus remained, for most institutions, an element of the workflow wholly dependent on contributions from externals. The number of selectors providing sites is typically small and their contributions are inevitably subjective. The resulting collections, whilst immensely valuable, are therefore mostly representative of the expertly selected sites and do not fully represent the sites frequently used in a more social setting.

Crowd sourcing is an opportunity to develop a new approach to this problem. It taps into the growth of social networks to outsource tasks typically performed by an employee or contractor, to an undefined, large group of people or community (a "crowd") (Wikipedia, 2011). It is a particularly attractive option in the current economic climate, where we are all being asked to 'do more with less'. A number of cultural heritage institutions and/or projects have already begun to leverage the power of the crowd for digitised collections, including the National Library of Australia (through Trove), the Transcribe Bentham project at ULCC, and the National Library of Finland (through the DigitalKoot program). No such projects have yet been launched for web archives.

This project will develop an automated approach to selection for web archiving based on the principles of crowd sourcing. It has been awarded partial funding from the International Internet Preservation Consortium and supports the forthcoming web archiving strategy by increasing the number of selections and automating the selection process.

**Links**

- [Twitter API](#)
- [UK Web Archive](#)

**Sponsors**

- [IIPC](#)

[Home](#)

Home | About

## TwitterVane

Crowd sourcing for Web Archiving

Collections Reports Streamed Tweets

**Collections**

Collection	Start Date	End Date	Search Terms	
News	January 29, 2013	February 28, 2013	news	Delete
#MyLifeIn5Words	January 29, 2013	February 13, 2013	#MyLifeIn5Words	Delete
Paul Scholes	January 29, 2013	February 13, 2013	Paul Scholes	Delete
#IReallyDislike	January 29, 2013	February 13, 2013	#IReallyDislike	Delete
#LyricsWeAllKnow	January 30, 2013	February 13, 2013	#LyricsWeAllKnow	Delete
#APhotoWith	January 30, 2013	February 13, 2013	#APhotoWith	Delete
Sir Paul Holmes	February 1, 2013	February 4, 2013	Paul Holmes Sir Paul Holmes #Paulholmes	Delete
Novopay	February 1, 2013	February 4, 2013	Novopay #Novopay #Novodump	Delete
eqnz	February 10, 2013	February 13, 2013	#eqnz	Delete
Sevens	February 1, 2013	February 4, 2013	#sevans #NZ7s	Delete
Iowa Senate Race	February 1, 2013	February 5, 2013	#iasen	Delete
Secretary Clinton leaves DoS	February 1, 2013	February 5, 2013	#hillary	Delete
testing a really long collection name to see what happens to layout of the collections table	February 1, 2013	February 2, 2013	#hilary	Delete
Oil Painting	February 4, 2013	February 5, 2013	Oil Painting techniques oil painting	Delete
Art History	February 3, 2013	February 6, 2013	Art history history of art	Delete
#etiquette	February 8, 2013	February 10, 2013	#etiquette #manners #advice	Delete
	February	February	#benedictovix	

**Links**

- [Twitter API](#)
- [UK Web Archive](#)

**Sponsors**

- [IIPC](#)



## Recent developments

- In the UK Non-print Legal Deposit & copyright law new exceptions
- Scholars are already interested in researching the web, and using web archives for research
- A new set of projects and initiatives: opening up web archives to utilise their potential research value
  - [NetLab](#)
  - [WebART: Web Archive Retrieval Tools](#) under the Continuous Access to Cultural Heritage (CATCH) programme
  - [Analytical Access to Dark Domain Archive](#) & [Big Data](#): Demonstrating the Value of the UK Web Domain Dataset for Social Science Research
  - [Digital Methods Initiative](#): providing training for next generation scholars & tools for internet research
  - Web Science 2013 call for papers includes a strand “Digital humanities, webarchiving techniques and scholarly uses of Web archives”

## 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 archives can be used for references as well as analytics
- Restricted access & technical limitations undermine the value of web archives but there is plenty we can do to bring web archives to the scholars
  - Highlight our USPs
  - Document what's missing
  - Fit in with researchers' workflow – how they do research
  - Full potential of web archives are yet to be exploited