

The Past Issue of the Web

*ACM Web Science Conference 2011
14-17 June Koblenz, Germany*

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

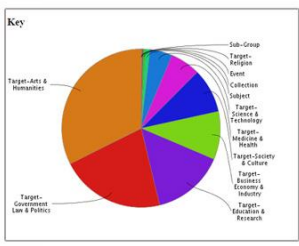
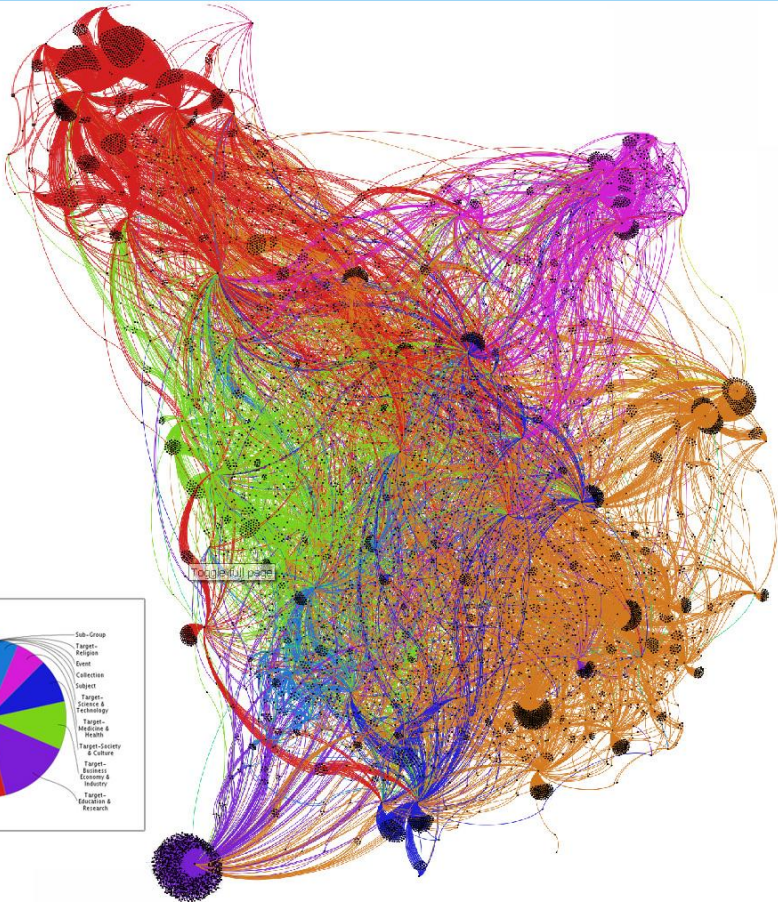
■ Who

- 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



Subject hierarchy visualisation [UK Web Archive](#)

- ~ 10,000 websites collected since 2004
- ~ 37,000 instances

- 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

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

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 search

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
Archived 23 Oct 2009	Archived 27 Apr 2010	Sorry, no thumbnail yet	Sorry, no thumbnail yet	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)

search

Advanced search

Your comments

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

Provided by:
BRITISH LIBRARY

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

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
- Legal deposit etc.
- Collection Care
- Press Room
- Contact us

News

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

Opens Fri 30 April

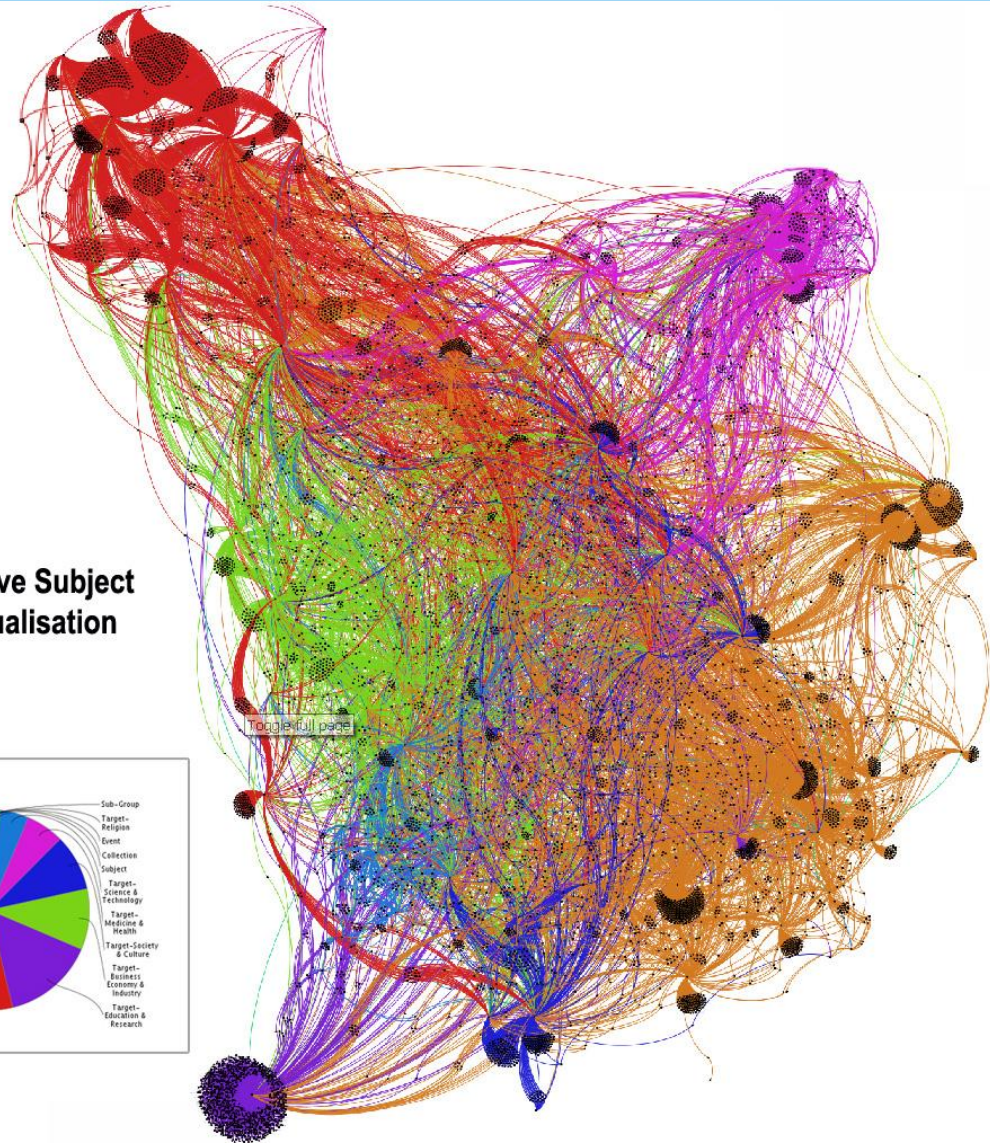
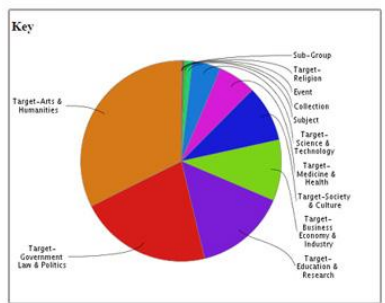
- Preview it online
- Read Curators' blog

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](#)

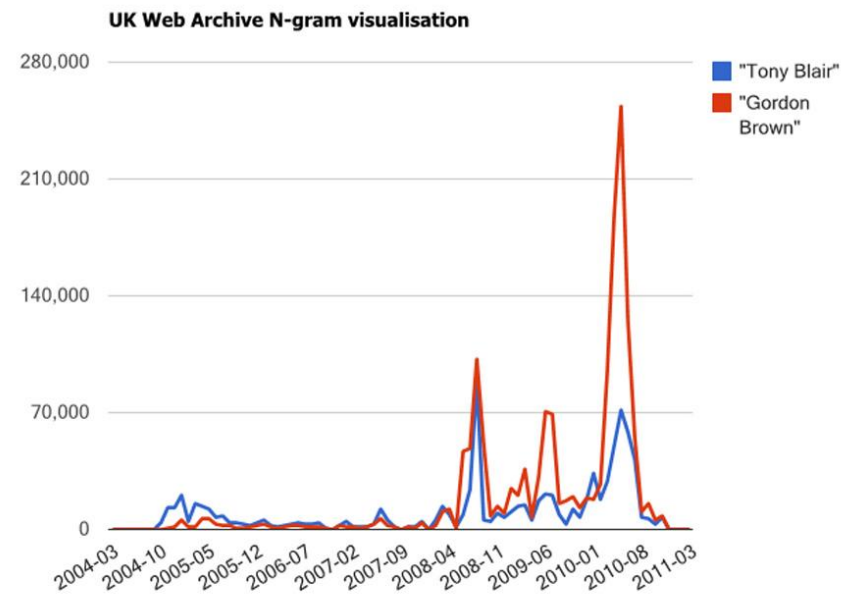
Quick search

Title (for a specific archived website)

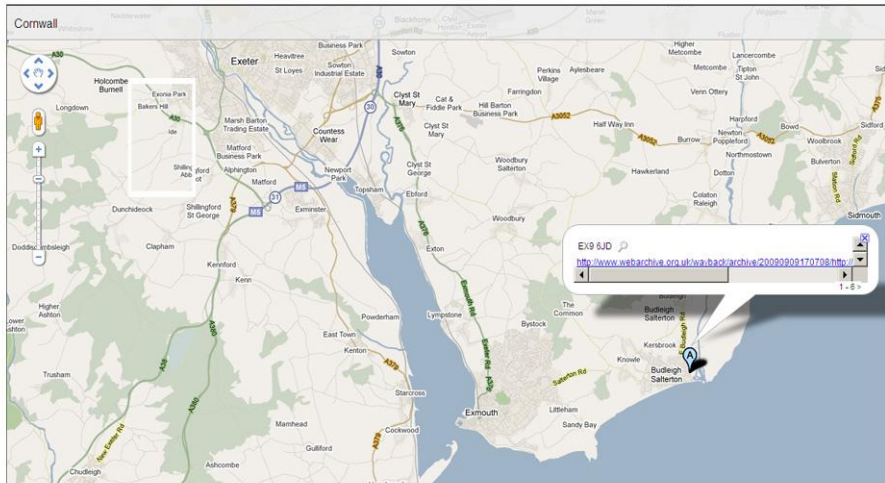
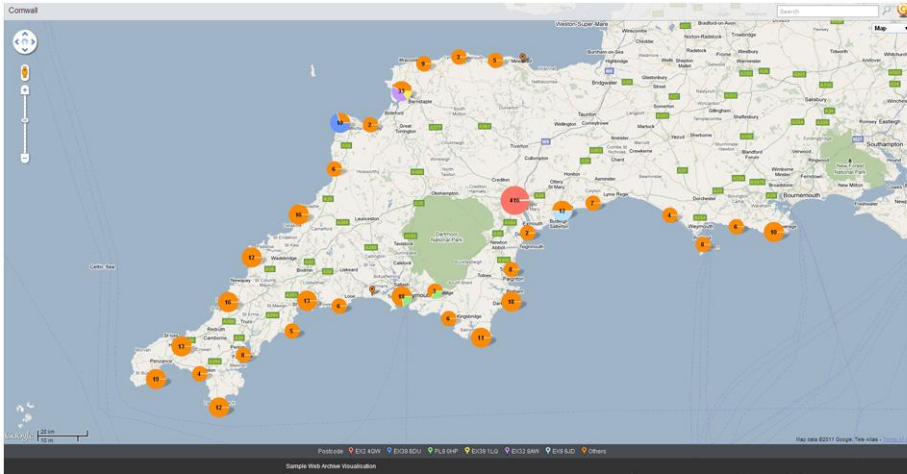
Full text (across all the archived websites)

[Advanced search](#)

UK Web Archive N-gram



The value of the haystacks – postcode-based access



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

UK WEB ARCHIVE

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.southdevonaoib.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) for ferry to South Sands from Whitestrand Pontoon. Pay and display. 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 Stationary Office © Crown copyright. Unauthorised reproduction infringes Crown copyright and may lead to prosecution or civil proceedings. Natural England Licence Number: 100046223. The Ordnance Survey mapping is included purely to provide a contextual backdrop for the walk and cannot be used for any other purpose.

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
- Web Science can help!
 - Already researching many of the issues
 - Use web archives to study the web