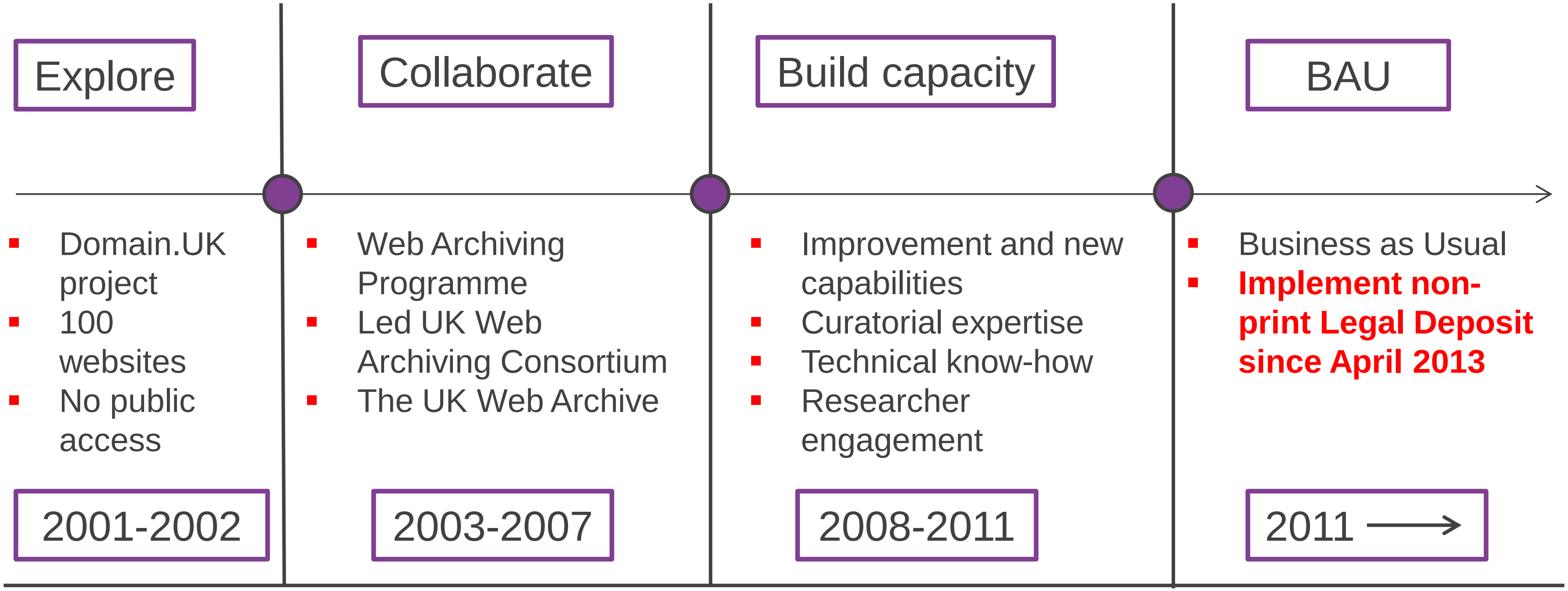
10 Years of archiving the UK Web

Helen Hockx-Yu

Head of Web Archiving, British Library

Web Archiving @ the British Library



Explore

- Domain.UK project
- 100 websites
- No public access

2001-2002

Collaborate

- Web Archiving Programme
- Led UK Web Archiving Consortium
- The UK Web Archive

2003-2007

Build capacity

- Improvement and new capabilities
- Curatorial expertise
- Technical know-how
- Researcher engagement

2008-2011

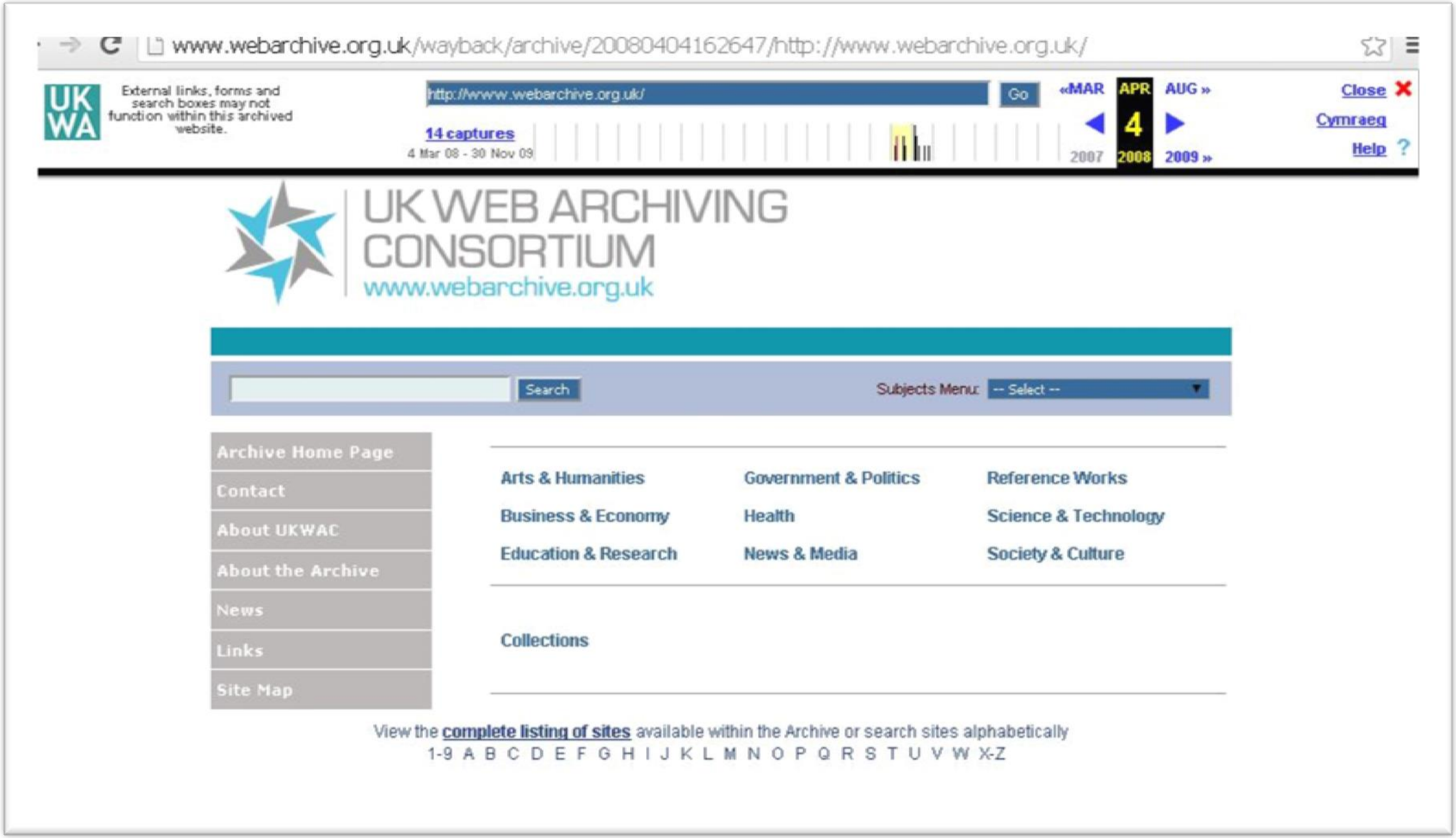
BAU

- Business as Usual
- **Implement non-print Legal Deposit since April 2013**

2011 →

Collect UK digital heritage and provide continued access to archived web resources

The UK Web Archiving Consortium



www.webarchive.org.uk/wayback/archive/20080404162647/http://www.webarchive.org.uk/

External links, forms and search boxes may not function within this archived website.

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

14 captures
4 Mar 08 - 30 Nov 09

UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk

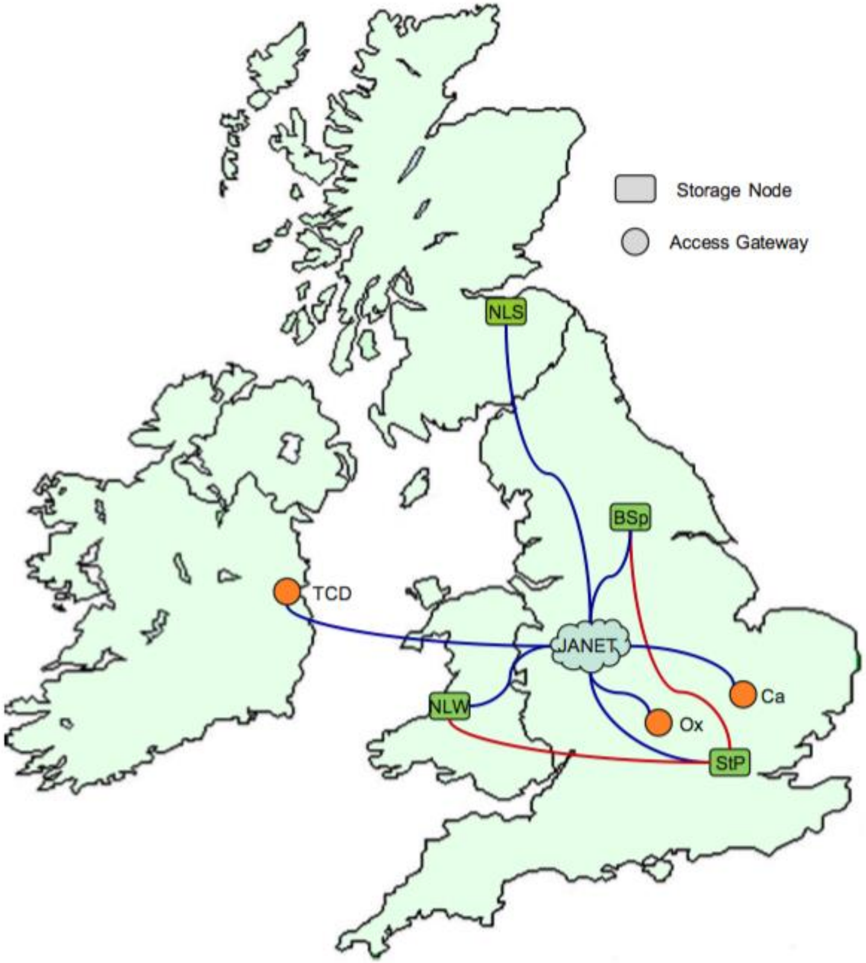
View the complete listing of sites available within the Archive or search sites alphabetically
1-9 A B C D E F G H I J K L M N O P Q R S T U V W X-Z

6th April 2013…

- Legal Deposit Libraries (Non-Print Works) Regulations 2013
- Extension of existing legal framework
- Systematic collection of UK's published output for heritage & preservation
- By 6 UK Legal Deposit Libraries

The Digital Library System

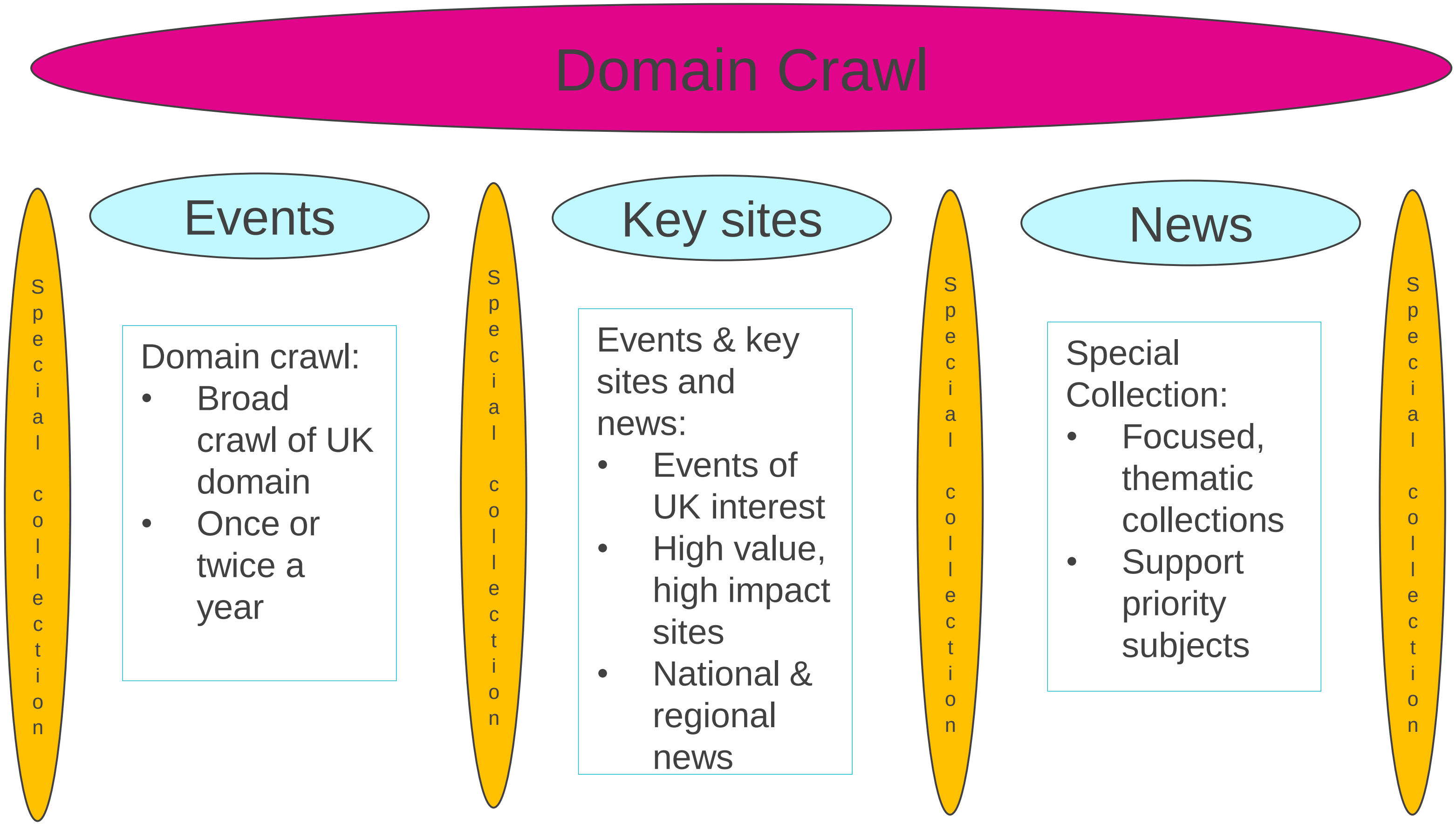
- 4 Nodes (Complete Copies)
  - British Library, St. Pancras
  - British Library, Boston Spa
  - National Library of Wales
  - National Library of Scotland
- Additional Access Points
  - Bodleian Library, Oxford
  - Cambridge University Library
  - Trinity College Library, Dublin



The UK Web Domain

- Over 10 million .uk registered domain
- UK organisations also use non .uk domain names (eg .com or .org)
    - Exact scale unknown
- Non-print Legal Deposit applies to
    - the open (freely available) web: .uk
    - other UK-published (non .uk) websites, such as .com, .org…
        - made available to the public by a person or an organisation and the activities relating to the creation or the publication of the work take place within the United Kingdom

Collecting strategy



Domain Crawl

Special collection

Events

Domain crawl:

- Broad crawl of UK domain
- Once or twice a year

Special collection

Key sites

Events & key sites and news:

- Events of UK interest
- High value, high impact sites
- National & regional news

Special collection

News

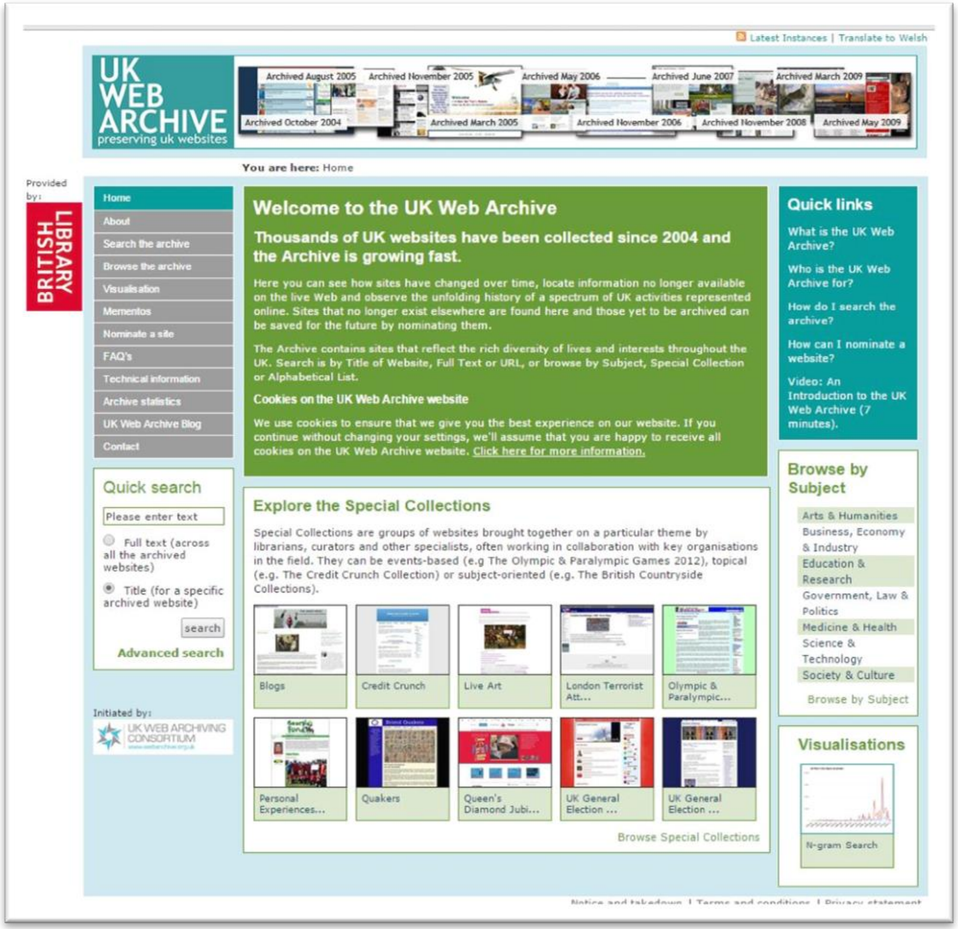
Special Collection:

- Focused, thematic collections
- Support priority subjects

Special collection

Open UK Web Archive

- Based on websites owners’ permissions
- 24/7 access
- ~68, 000 point-in-time snapshots of over 15,000 selected UK websites
- 21TB
- Continues to grow
- Going forward: showcase the Legal Deposit content



Welcome to the UK Web Archive

Thousands of UK websites have been collected since 2004 and the Archive is growing fast.

Here you can see how sites have changed over time, locate information no longer available on the live Web and observe the unfolding history of a spectrum of UK activities represented online. Sites that no longer exist elsewhere are found here and those yet to be archived can be saved for the future by nominating them.

The Archive contains sites that reflect the rich diversity of lives and interests throughout the UK. Search is by Title of Website, Full Text or URL, or browse by Subject, Special Collection or Alphabetical List.

Cookies on the UK Web Archive website

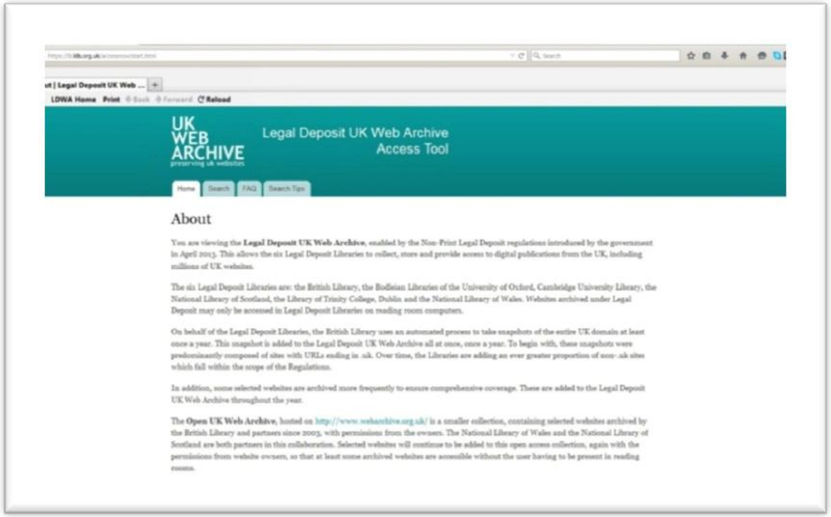
We use cookies to ensure that we give you the best experience on our website. If you continue without changing your settings, we'll assume that you are happy to receive all cookies on the UK Web Archive website. Click here for more information.

Explore the Special Collections

Special Collections are groups of websites brought together on a particular theme by librarians, curators and other specialists, often working in collaboration with key organisations in the field. They can be 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).

Legal Deposit UK Web Archive

- Annual UK Domain Crawl
  - 2013: 31TB, 1.6 billion URLs, 4 million hosts
  - 2014: 57TB including 2.5 non .uk domains
- Events, news sites, key sites
  - ~15TB
- Discoverable through British Library's online catalogue
- Full-text searchable in the reading rooms



About

You are viewing the **Legal Deposit UK Web Archive**, enabled by the Non-Print Legal Deposit regulations introduced by the government in April 2013. This allows the six Legal Deposit Libraries to collect, store and provide access to digital publications from the UK, including millions of UK websites.

The six Legal Deposit Libraries are: the British Library, the Bodleian Libraries of the University of Oxford, Cambridge University Library, the National Library of Scotland, the Library of Trinity College, Dublin and the National Library of Wales. Websites archived under Legal Deposit may only be accessed in Legal Deposit Libraries on reading room computers.

On behalf of the Legal Deposit Libraries, the British Library uses an automated process to take snapshots of the entire UK domain at least once a year. This snapshot is added to the Legal Deposit UK Web Archive all at once, once a year. To begin with, these snapshots were predominantly composed of sites with URLs ending in .uk. Over time, the Libraries are adding an ever greater proportion of non-.uk sites which fall within the scope of the Regulations.

In addition, some selected websites are archived more frequently to ensure comprehensive coverage. These are added to the Legal Deposit UK Web Archive throughout the year.

The **Open UK Web Archive**, hosted on http://www.webarchive.org.uk/ is a smaller collection, containing selected websites archived by the British Library and partners since 2003, with permissions from the owners. The National Library of Wales and the National Library of Scotland are both partners in this collaboration. Selected websites will continue to be added to this open access collection, again with the permissions from website owners, so that at least some archived websites are accessible without the user having to be present in reading rooms.

Same job, just bigger and more complex

"Hockx-Yu is now on the front line of the most ambitious expansion of the British Library's archiving capability in more than 300 years. At the stroke of midnight on April 5, 2013, legislation known as the Legal Deposit regulations came into force, charging the library with capturing the contents of the entire U.K. web domain—every site carrying the .uk suffix—preserving the material and making it publicly accessible."

Jacobson (2014)

TECH & SCIENCE

Inside the Struggle to Preserve the World's Data

BY PHILIP JACOBSON / JULY 2, 2014 12:21 PM EDT

Newsweek

" the British Library's web archiving operation, [is] rated by experts in the field as among the best in the business. "

Web archiving: the global picture

- Survey of web archiving initiatives (Daniel Gomes et al 2010)
    - 42 web archiving initiatives across 26 countries since 1996
    - 6.6PB archived web resources
- International Internet Preservation Consortium (IIPC) - from 12 founding members in 2003 to nearly 50 in 2015
    - Most members are national memory institutions
- How much of the Web is Archived (Scott Ainsworth et al, 2012)

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

- BL and CDNL (association of chief executives of national libraries) survey: 30 nations have legislations by 2012 enabling web harvesting

Types of web archives

- Global: Internet Archive’s Wayback Machine
- National: UK, Denmark, France, Finland and many other countries
- Sub-set of a national domain: eg UK Government Web Archive, Canadian Government Web Archive
- Topical collections: eg Human rights
- Collaborative: Archive–It service by Internet Archive
- Corporate: a single organisation’s web estate, eg UK Parliament Web Archive

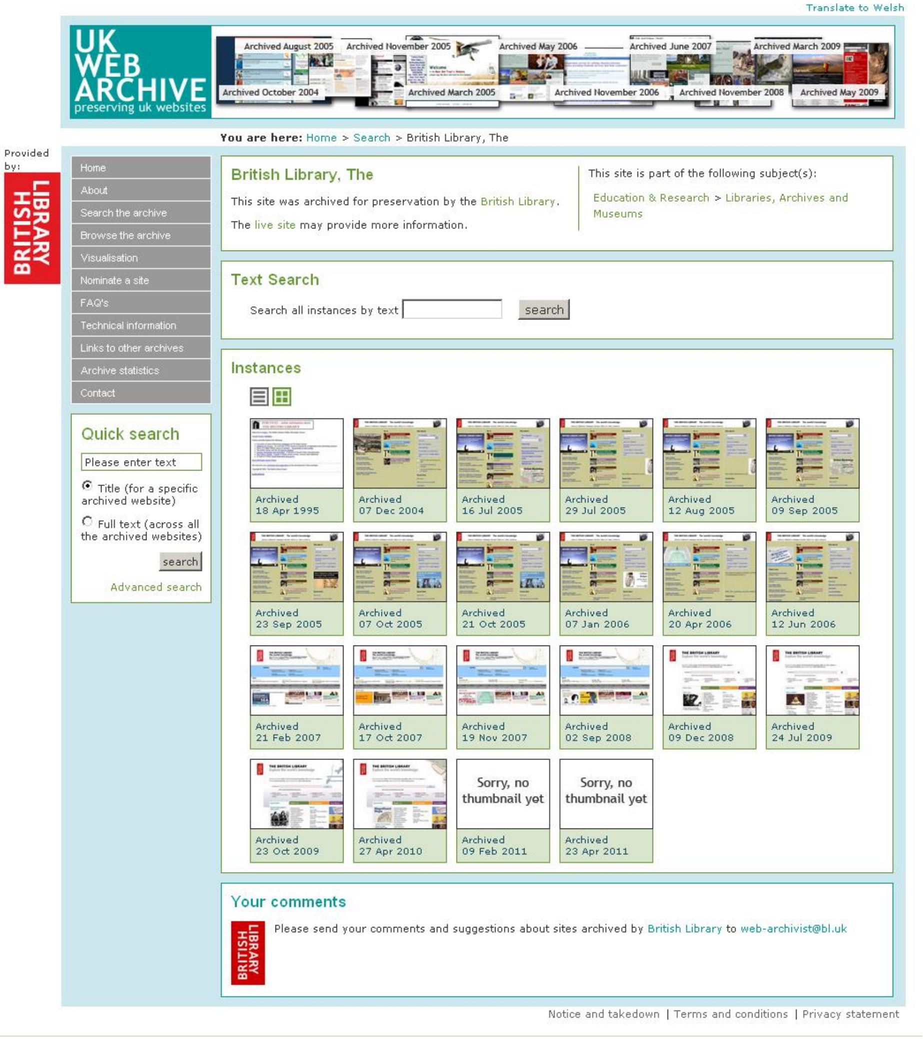
Reflections / observations

- High concentration of use of technology and practice
- Modern browsers can render virtually any webpage. But the same pages often defeat our purpose-built crawlers
- National approach to archiving the global web
- Does size matter?
- Role of Legal Deposit
- Library/Archive discourse
- What is driving web archiving?

Access and use of web archives

- Completeness versus openness
  - Many countries don’t have Legal Deposit
  - Legal Deposit national collections have restricted access
- Not enough versus too much
  - Pre-selected or defined collections not relevant to all researchers; difficulty in finding relevant content in large scale web archive
- Arbitrary (national) boundaries often irrelevant to research question
- (Single) envisaged use case based on browsing of individual websites page by page

An archived website



Translate to Welsh

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

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

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 [ ] search

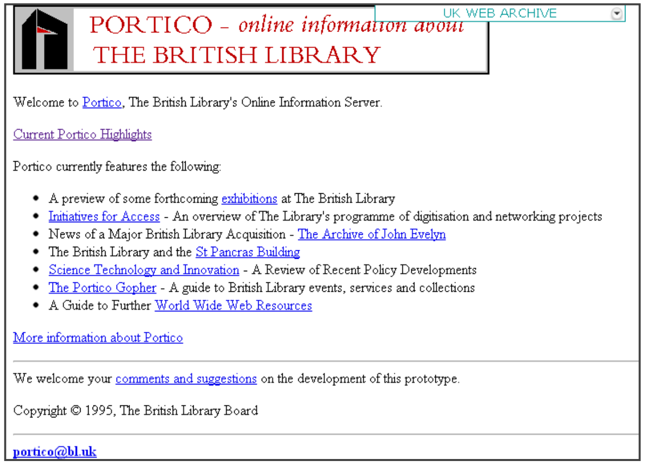
Instances

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 Archived 09 Feb 2011	Sorry, no thumbnail yet Archived 23 Apr 2011		

Your comments

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

Notice and takedown | Terms and conditions | Privacy statement



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:

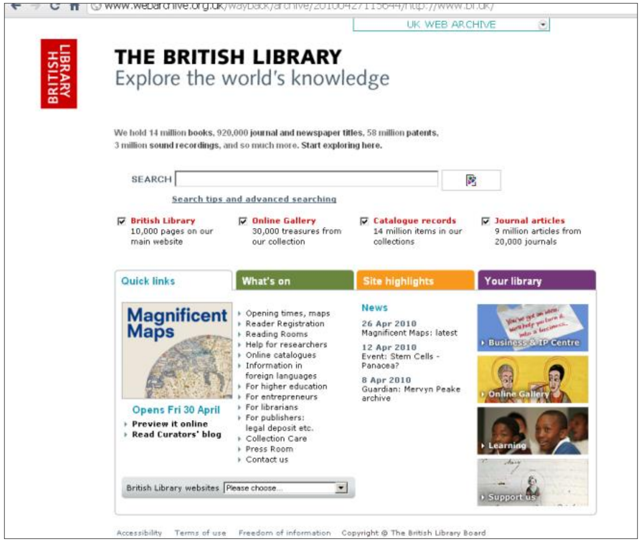
- A preview of some forthcoming exhibitions at The British Library
- Initiatives for Access - An overview of The Library's programme of digitisation and networking projects
- News of a Major British Library Acquisition - The Archive of John Evelyn
- The British Library and the St Pancras Building
- Science 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



www.webarchive.org.uk/wayback/archive/20100427115644/http://www.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

☑ **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

Opens Fri 30 April

- **Preview it online**
- **Read Curators' blog**

- 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?

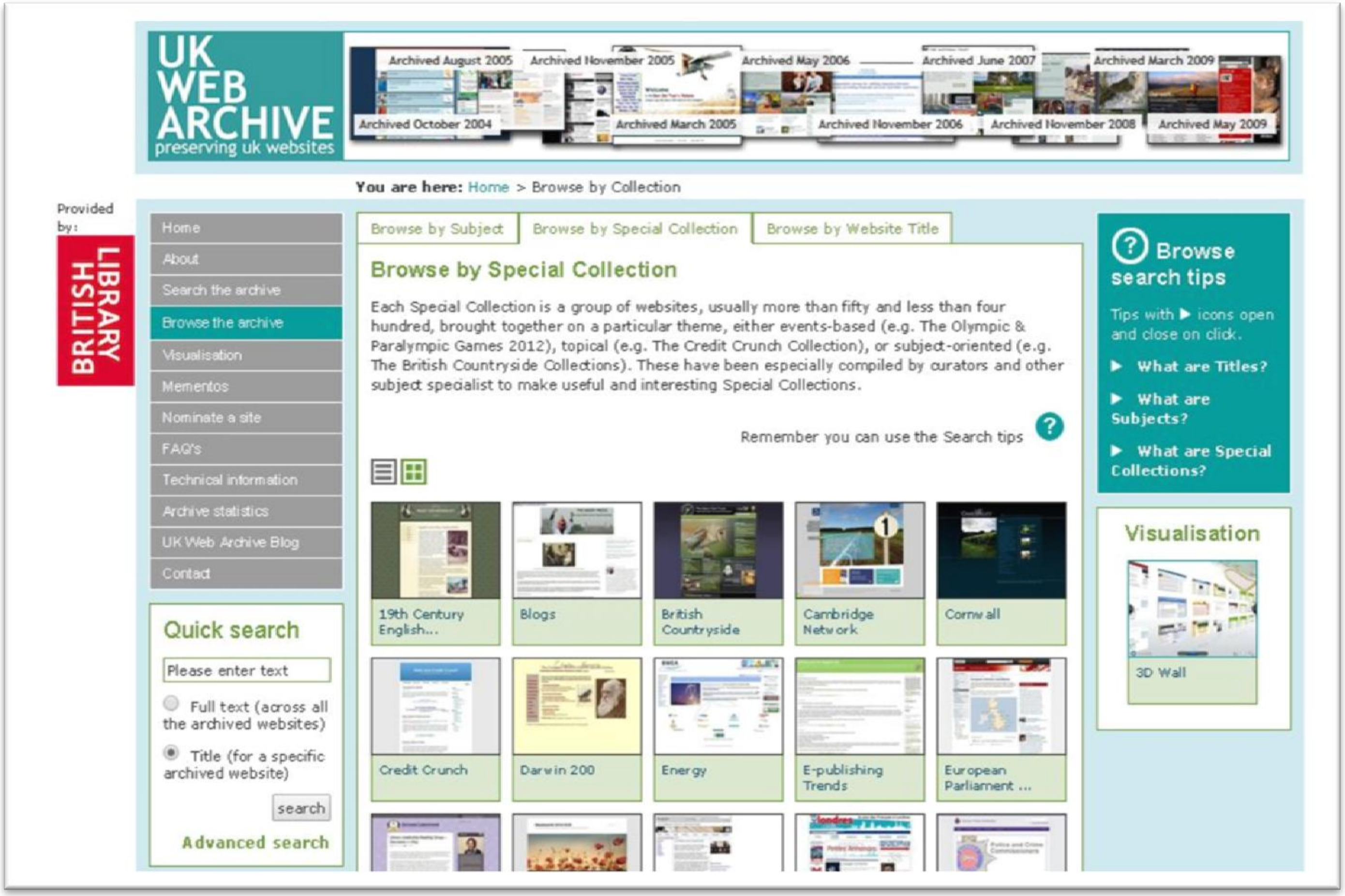
8 Apr 2010
Guardian: Mervyn Peake archive

- Business & IP Centre
- Online Gallery
- Learning
- Support us

British Library websites [Please choose...]

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

Collections



You are here: Home > Browse by Collection

- 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

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
- Cambridge Network
- Cornwall
- Credit Crunch
- Darwin 200
- Energy
- E-publishing Trends
- European Parliament ...

Browse search tips

Tips with ▶ icons open and close on click.

▶ What are Titles?

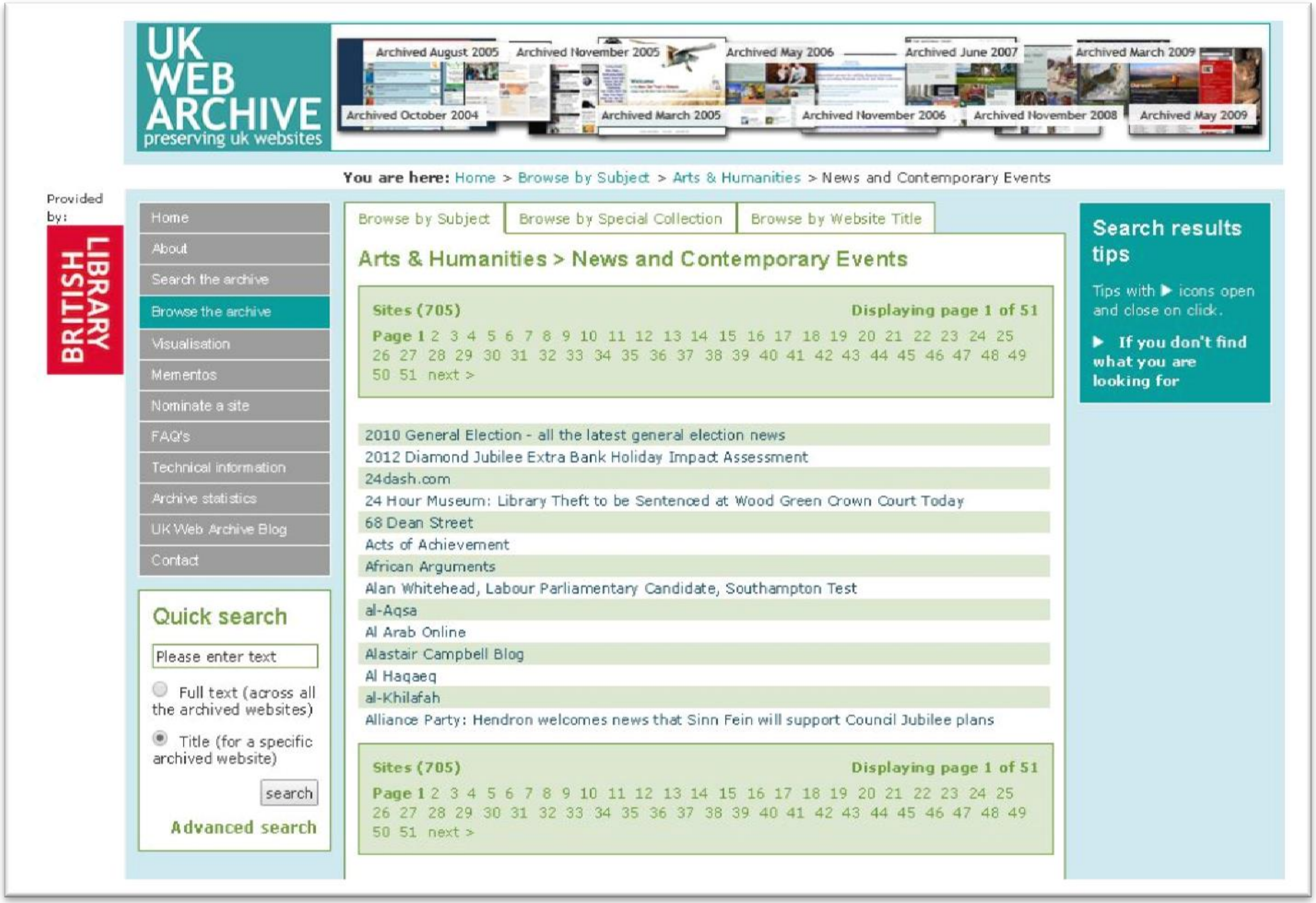
▶ What are Subjects?

▶ What are Special Collections?

Visualisation

3D Wall

Subjects



You are here: Home > Browse by Subject > Arts & Humanities > News and Contemporary Events

Arts & Humanities > News and Contemporary Events

Sites (705) **Displaying page 1 of 51**

Page 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 next >

2010 General Election - all the latest general election news
2012 Diamond Jubilee Extra Bank Holiday Impact Assessment
24dash.com
24 Hour Museum: Library Theft to be Sentenced at Wood Green Crown Court Today
68 Dean Street
Acts of Achievement
African Arguments
Alan Whitehead, Labour Parliamentary Candidate, Southampton Test
al-Aqsa
Al Arab Online
Alastair Campbell Blog
Al Haqaeq
al-Khilafah
Alliance Party: Hendron welcomes news that Sinn Fein will support Council Jubilee plans

Sites (705) **Displaying page 1 of 51**

Page 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 next >

Search results tips

Tips with ▶ icons open and close on click.

▶ **If you don't find what you are looking for**

Quick search

Please enter text

○ Full text (across all the archived websites)

● Title (for a specific archived website)

search

Advanced search

Legal Deposit UK Web Archive

- Full-text search, results can be refined by facets
  - Content type (html, pdf…)
  - Crawl year
  - (individual) Domain (bbc.co.uk; newsforscotland.com…)
  - Domain suffix (.co.uk; .org.uk; .ac.uk; .com)
  - Author (based on metadata extracted from web pages)
  - …
- Still has “collections” – selected and curated by curators, usually crawled within a fixed period of time, relating to an event
- Subjects information – added to selected sites only – included in the index

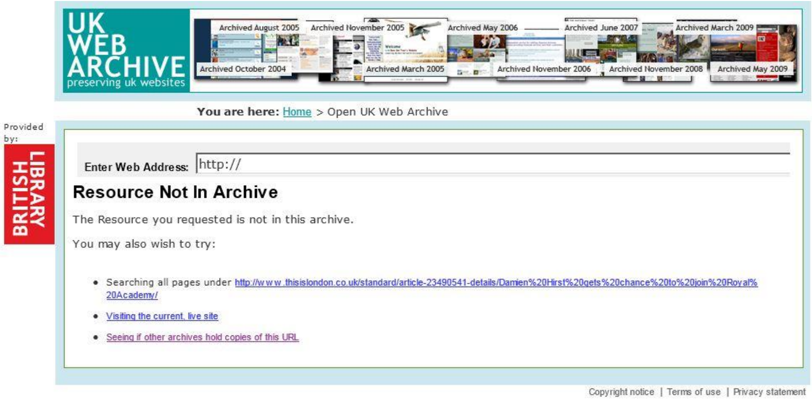
Issues of “document-centric” approach

- Rise of “digital scholarship”, taking advantage of the possibilities offered by technology
- Primacy of “text” as object of study no longer exists
- “Paratexts” play a crucial role in textual coherence of a website: header & footer drop-down menu, site map, breadcrumb, etc. (Brügger, 2010)
- “Distant reading” focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems (Moretti, 2000)

“Resource Not in Archive”

- Common error message but appears for different reasons

- Intended boundary
  - No permission for linked content
  - Not allow by robots.txt
  - Edge of a archive
  - Data limitation



You are here: Home > Open UK Web Archive

Enter Web Address: http://

Resource Not In Archive

The Resource you requested is not in this archive.

You may also wish to try:

- Searching all pages under http://w w w .thisislondon.co.uk/standard/article-23490541-details/Damien%20Hirst%20gets%20chance%20to%20join%20Royal%20Academy/
- Visiting the current, live site
- Seeing if other archives hold copies of this URL

Copyright notice | Terms of use | Privacy statement

- Technical limitations, e.g. dynamic content the crawler could not collect

- In UK Web Archive: avoid dead end in navigation
  - Search
  - Link to live web
  - Find archived copies elsewhere

There is so much more…

- Statistical overview, scale and distribution of a web national domain
- Evolution, pattern of change over time, eg domain names
- Space: geo location, postcodes
- Type of content, eg file format, language
- Correlation, eg between certain term and historical event
- Structure, linked entities and networks
- Other web archives

JISC UK Web Domain dataset (1996-2013)

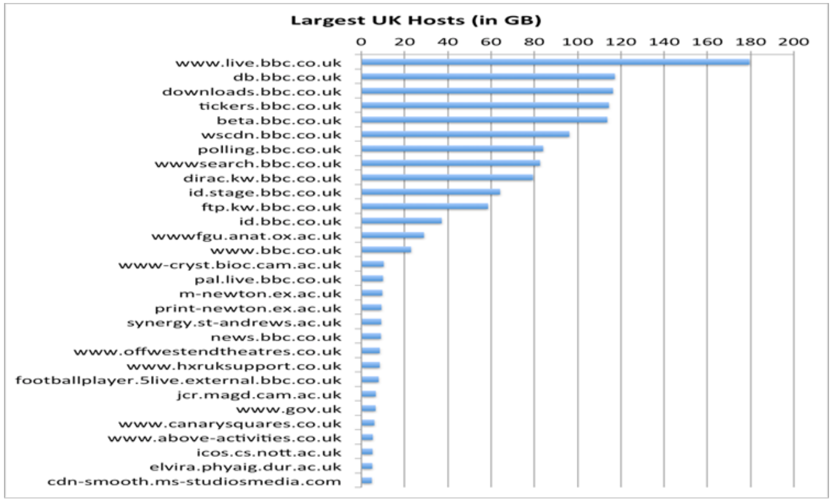
- Copies of UK websites extracted from the Internet Archives collection
  - 57TB, 4 billion URLs

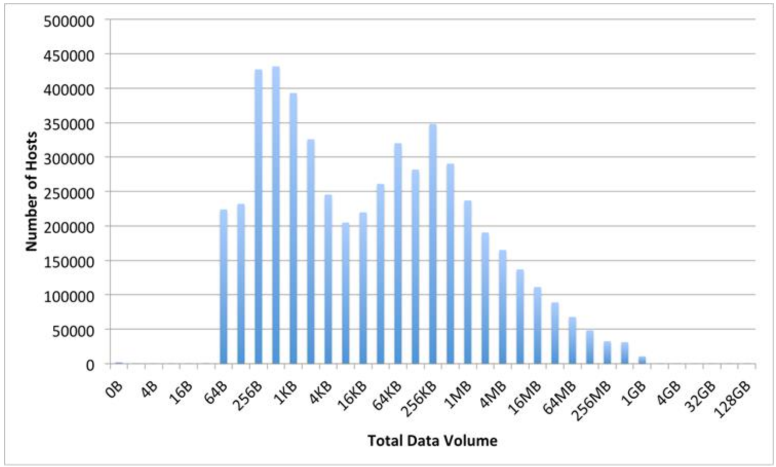
- Research agreement between JISC and IA, upholding IA's Terms of Use
  - Access via IA's Wayback Machine
  - Allows replication / extraction of secondary datasets: Format profile, Geoindex, Host Link Graph

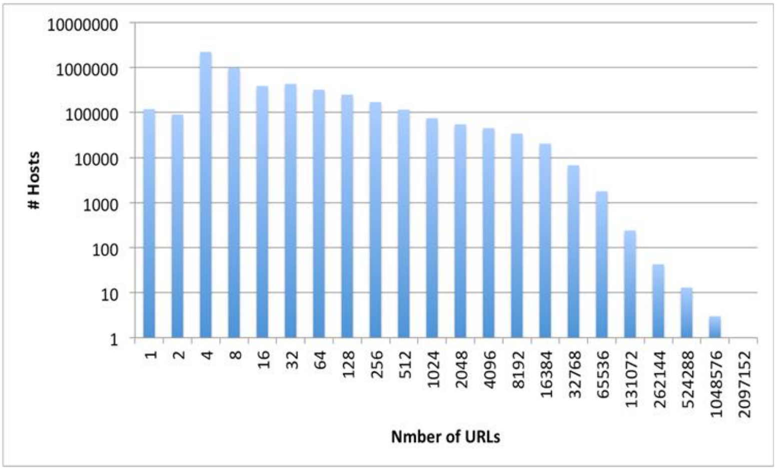
- BL hosts the dataset on behalf of JISC

- Serves as research test bed and used by various research projects
  - JISC funded: Analytical Access to the Domain Dark Archive (AADDA); Big data for political science
  - AHRC funded: Big UK Domain Data for Arts and Humanities
  - ERC funded: Alexandria

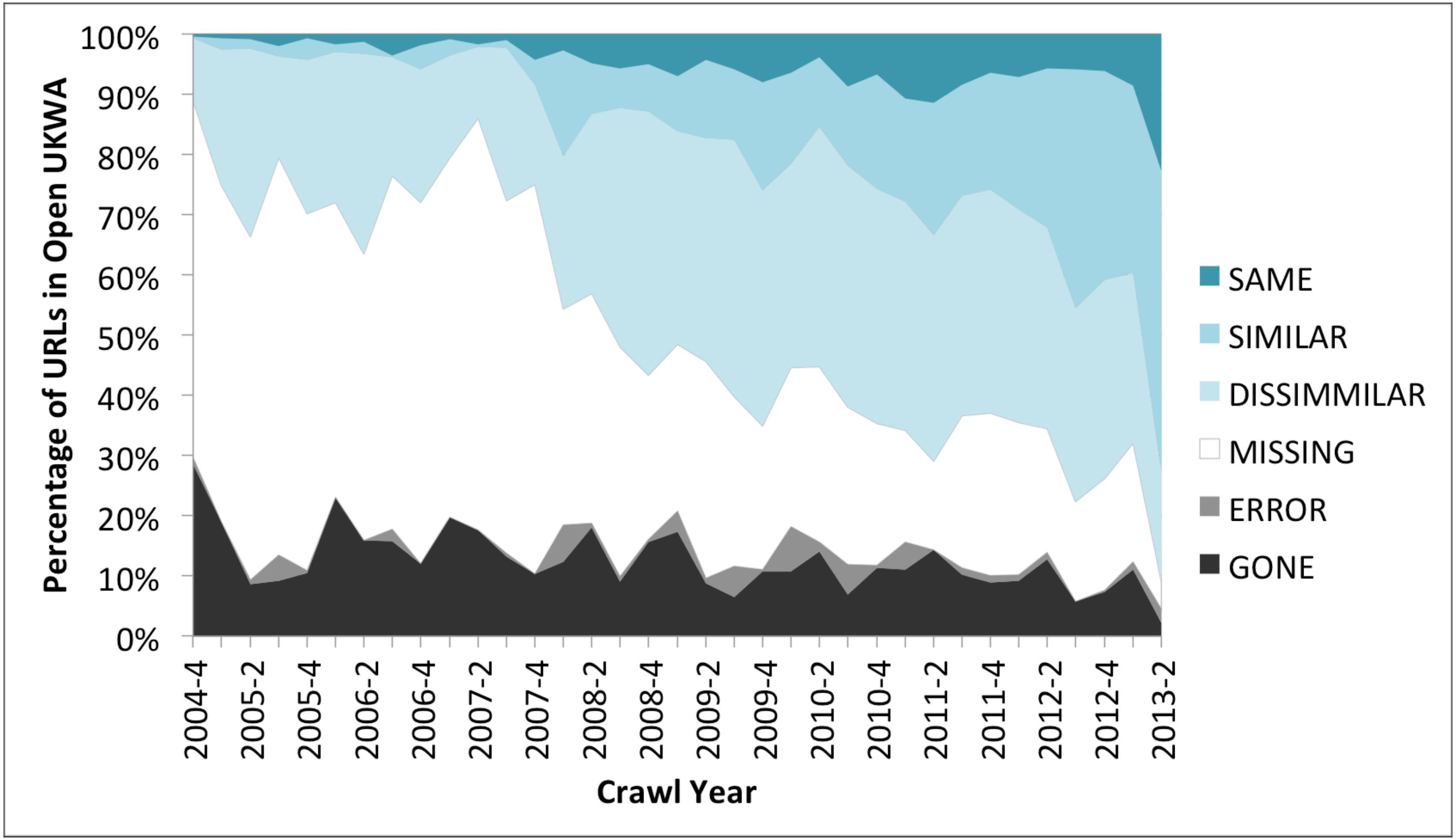
The UK Web in 2013



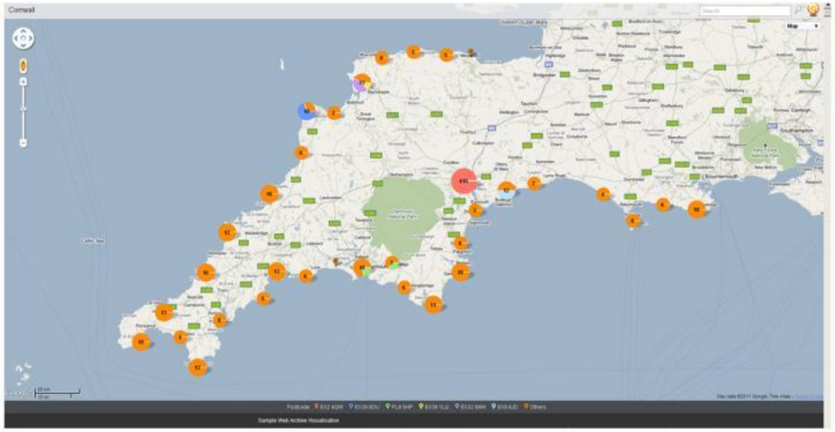


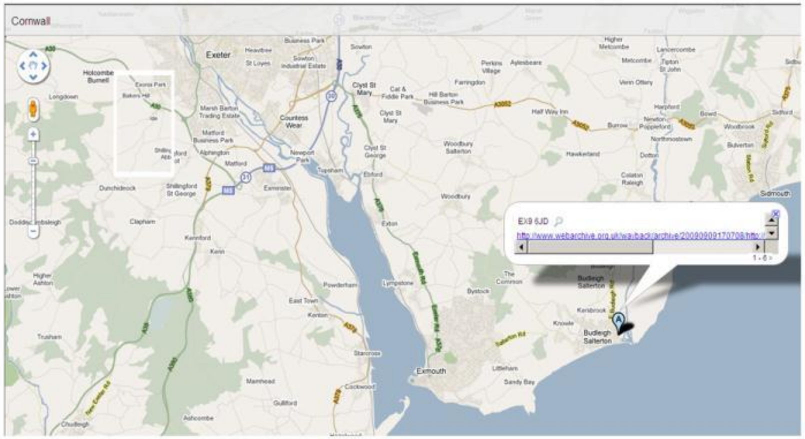


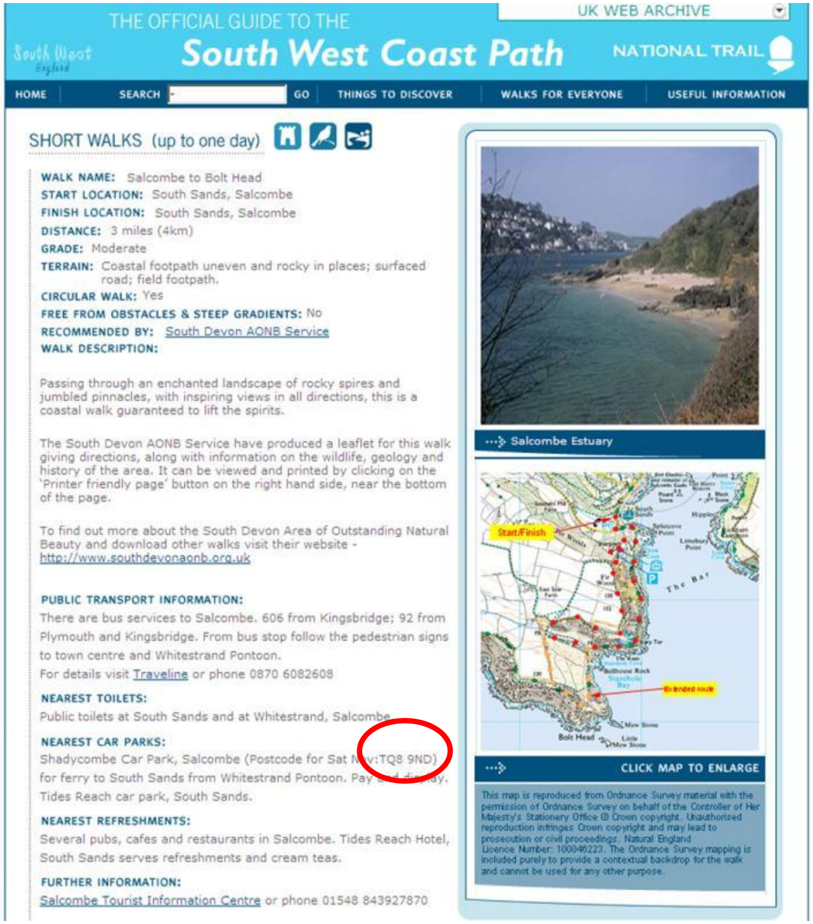
Evolution of the UK web (2004 -2013)



Postcode-based access




Cornwall
Exeter
Exmouth
Budleigh Salterton
Sidmouth
EX9 6JD
http://www.webarchive.org.uk/wayback/archive/20090909170708/http://
1 - 6 ›



THE OFFICIAL GUIDE TO THE

South West Coast Path

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) 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 Stationery 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.

HTML version analysis

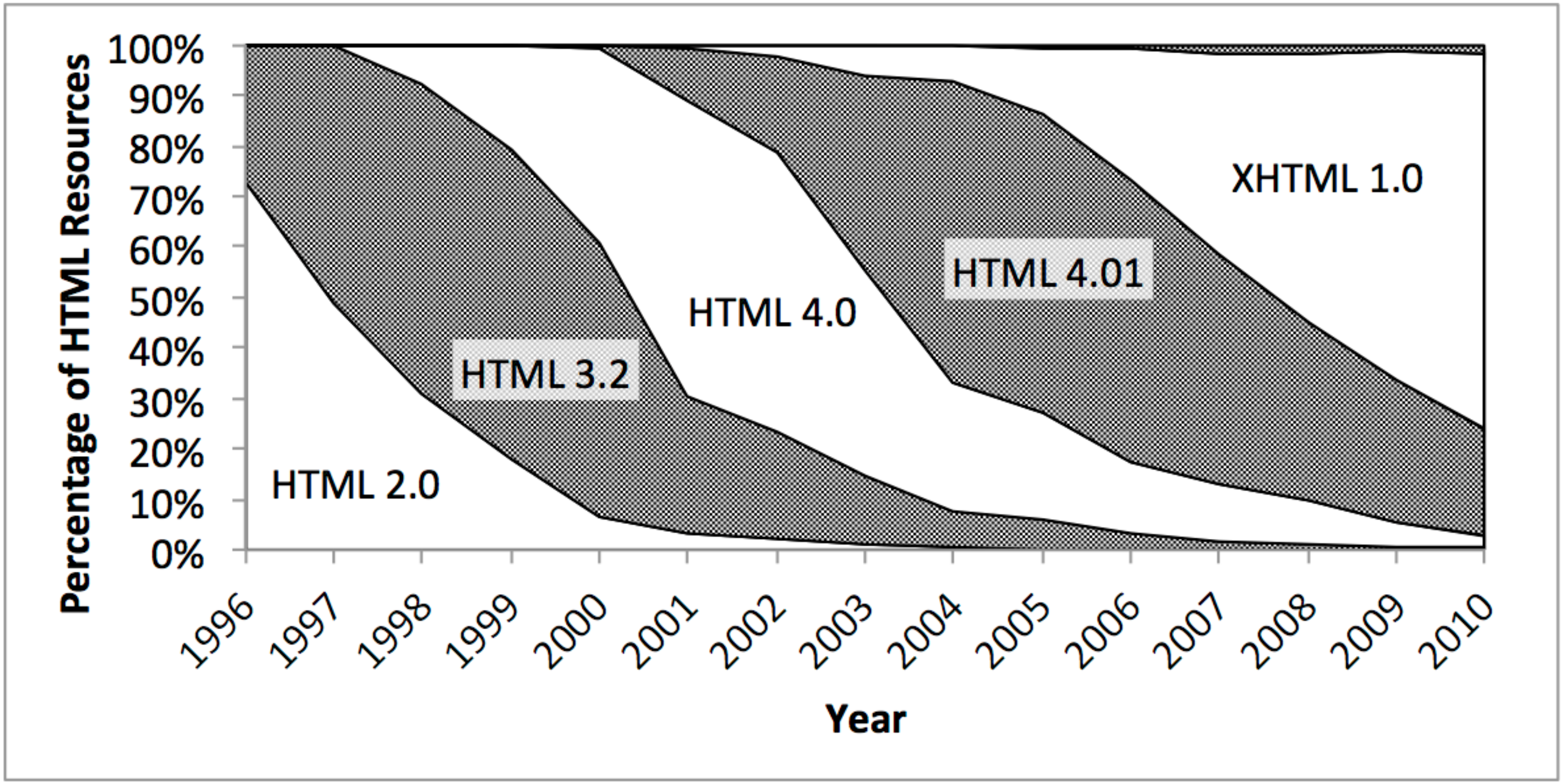
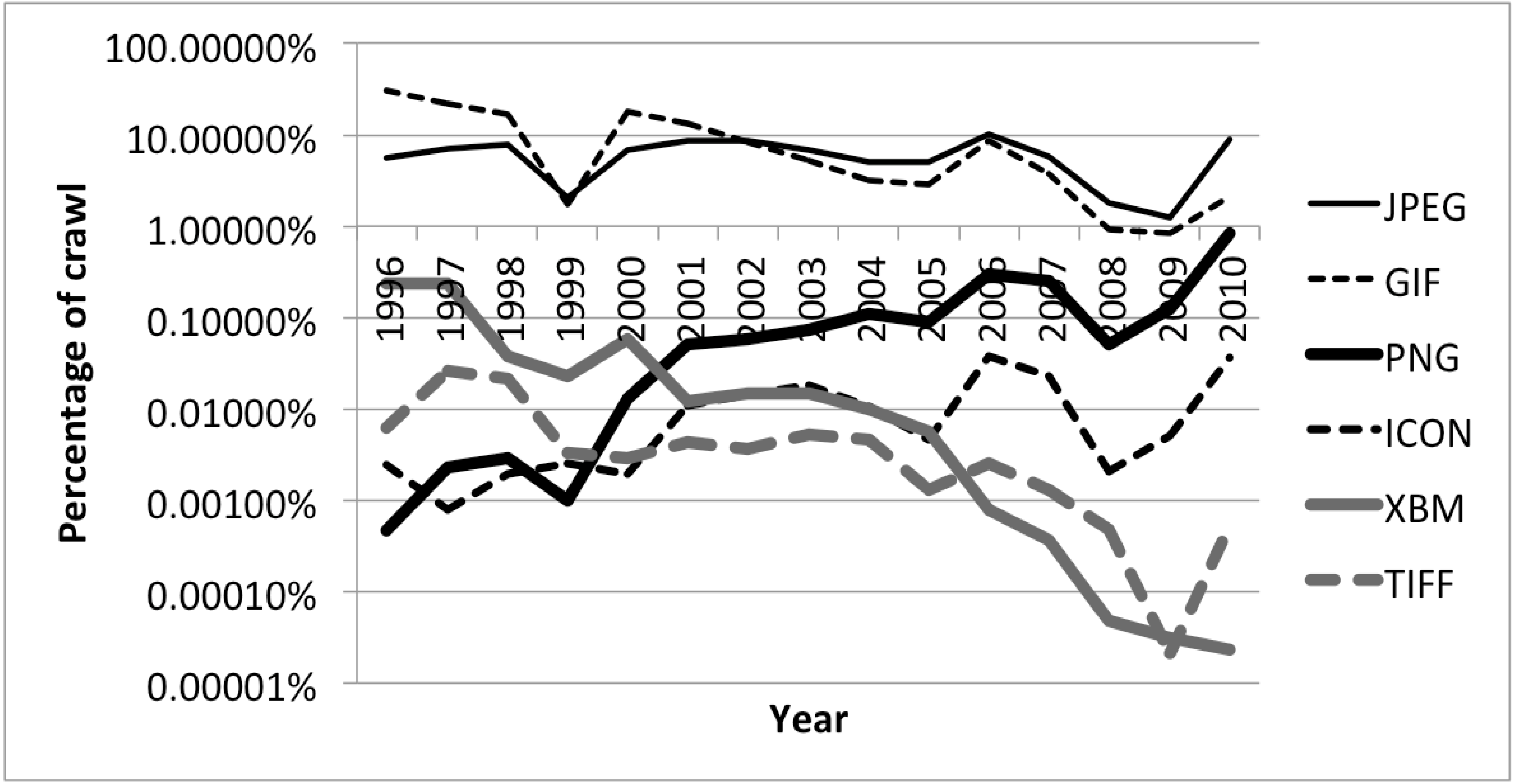
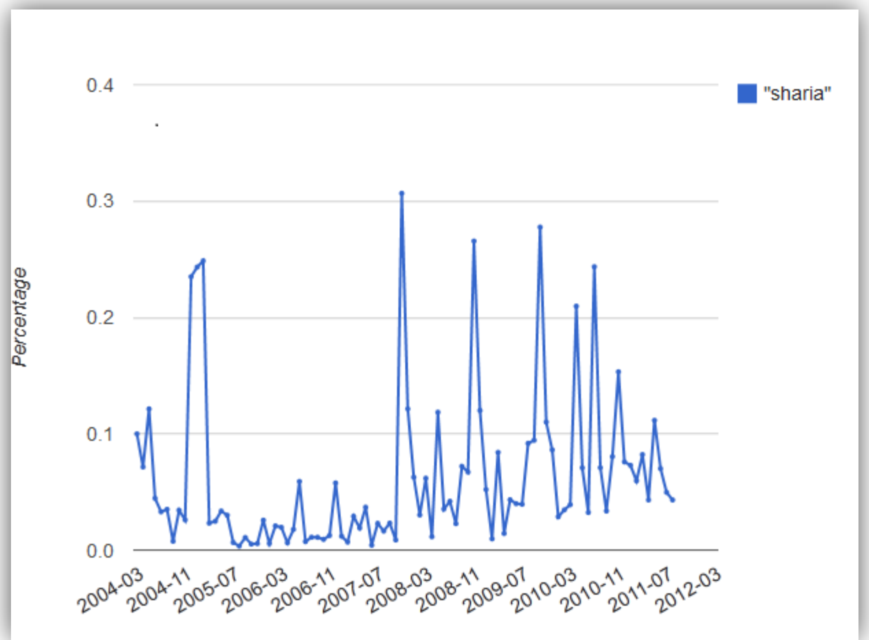
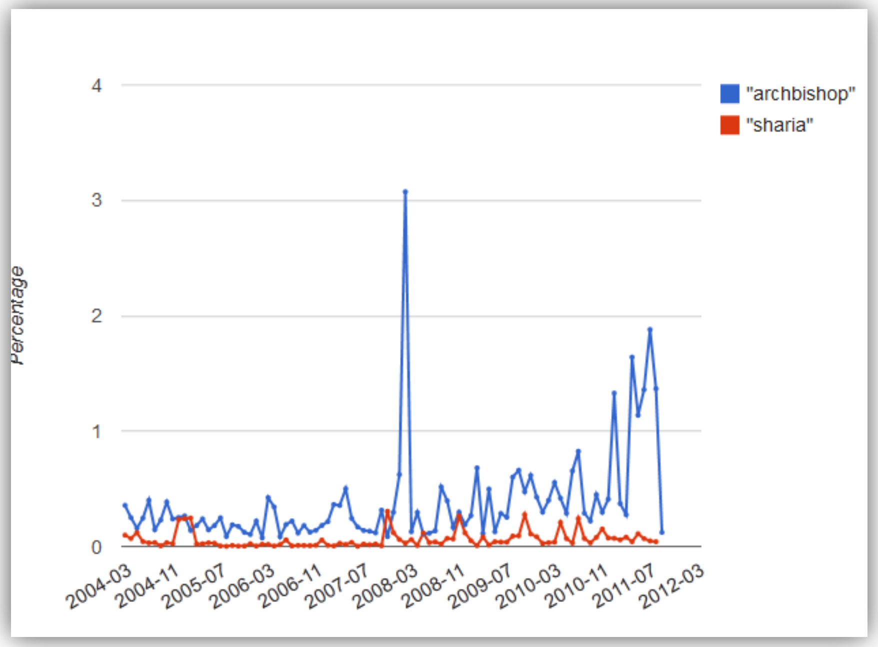


Image Format Analysis



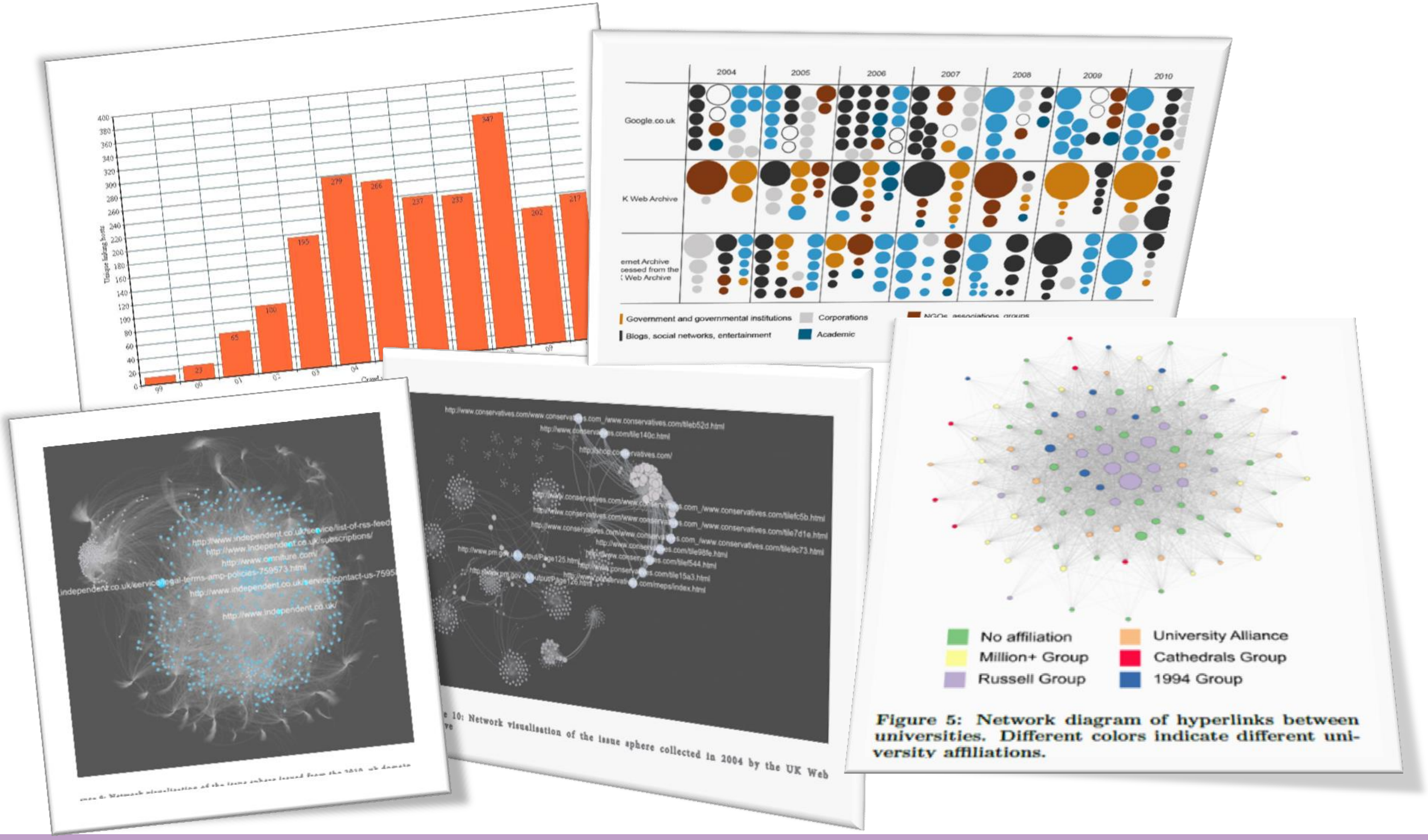
Using N-gram for scholarly research





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

Exploring links


400
380
360
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
Unique linking hosts
23
65
100
195
279
266
237
233
347
202
217
99
00
01
02
03
04
09
Crawl
2004
2005
2006
2007
2008
2009
2010
Google.co.uk
K Web Archive
ernet Archive
cessed from the
(Web Archive
Government and governmental institutions
Corporations
NGOs, associations, groups
Blogs, social networks, entertainment
Academic
http://www.independent.co.uk/service/list-of-rss-feeds
http://www.independent.co.uk/subscriptions/
http://www.omniture.com/
independent.co.uk/service/legal-terms-amp-policies-759573.html
http://www.independent.co.uk/service/contact-us-7595
http://www.independent.co.uk/
http://www.conservatives.com/www.conservatives.com_/www.conservatives.com/tileb52d.html
http://www.conservatives.com/tile140c.html
http://shop.conservatives.com/
http://www.conservatives.com/www.conservatives.com_/www.conservatives.com/tilefc5b.html
http://www.conservatives.com/www.conservatives.com_/www.conservatives.com/tile7d1e.html
http://www.conservatives.com/www.conservatives.com_/www.conservatives.com/tile9c73.html
http://www.conservatives.com/tile98fe.html
http://www.pm.gov.uk/output/Page125.html
http://www.conservatives.com/tilef544.html
http://www.pm.gov.uk/output/Page126.html
http://www.conservatives.com/tile15a3.html
http://www.conservatives.com/meps/index.html
No affiliation
Million+ Group
Russell Group
University Alliance
Cathedrals Group
1994 Group

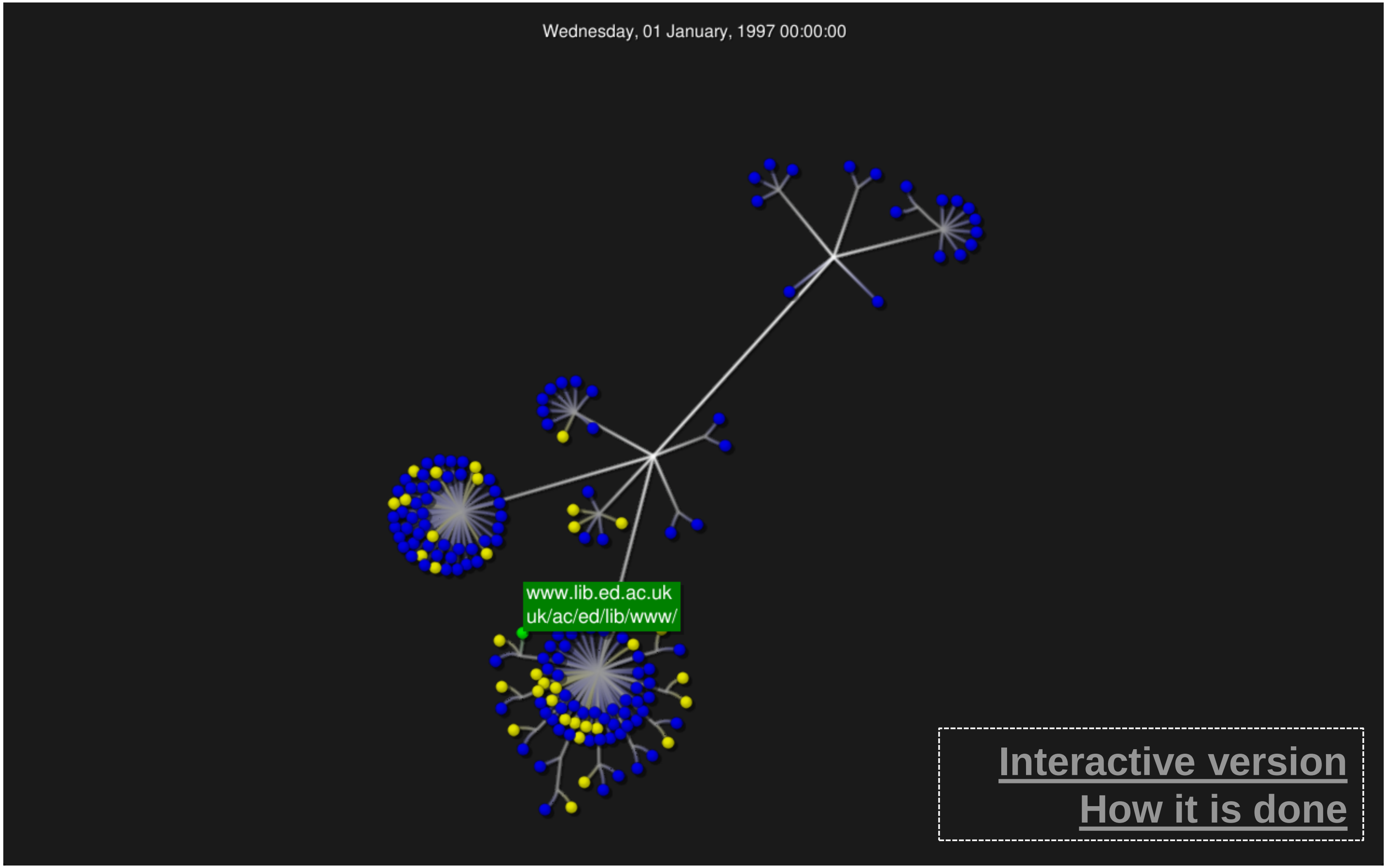
...ure 6: Network visualisation of the issue sphere issued from the 2010 .uk domain

...e 10: Network visualisation of the issue sphere collected in 2004 by the UK Web ...ve

Figure 5: Network diagram of hyperlinks between universities. Different colors indicate different university affiliations.

Courtesy of Peter Webster, Rainer Simon and Jules Mataly

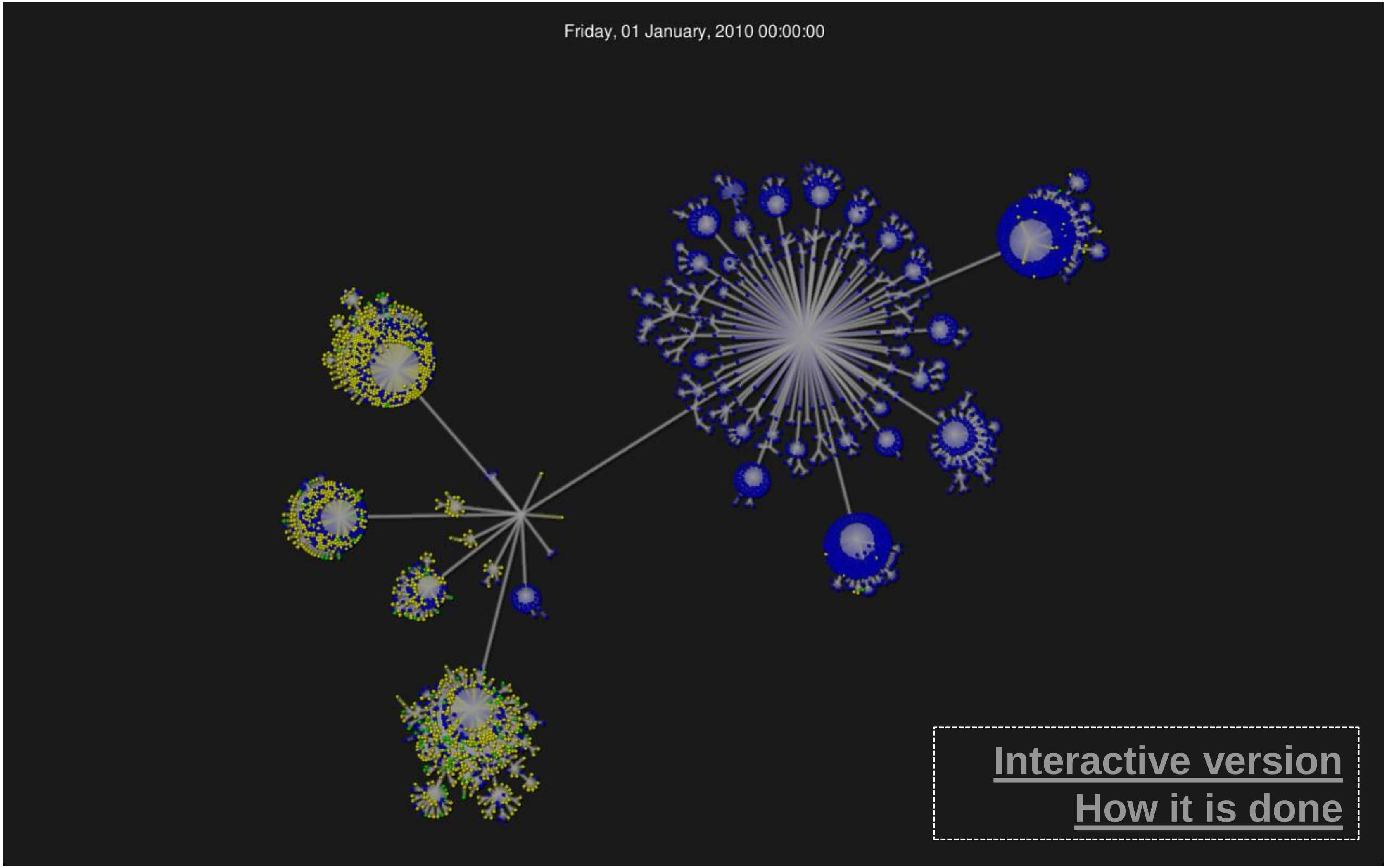
Visualising links (to and from bl.uk 1996)



[Interactive version](#)

[How it is done](#)

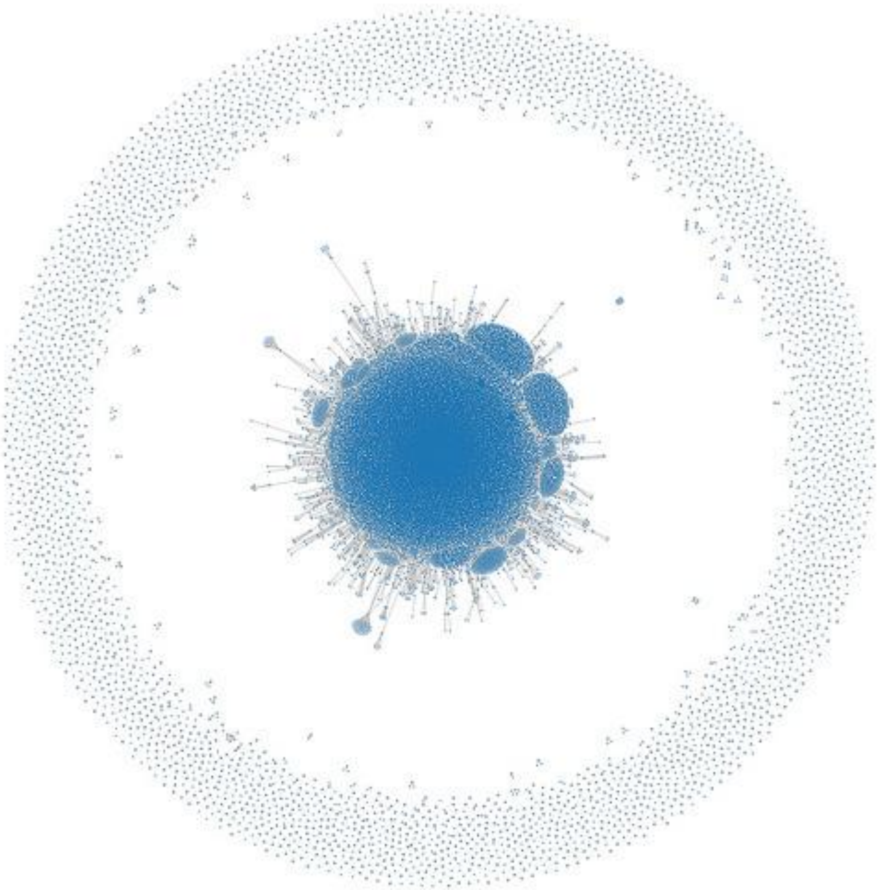
Visualising links (to and from bl.uk 2010)



Interactive version

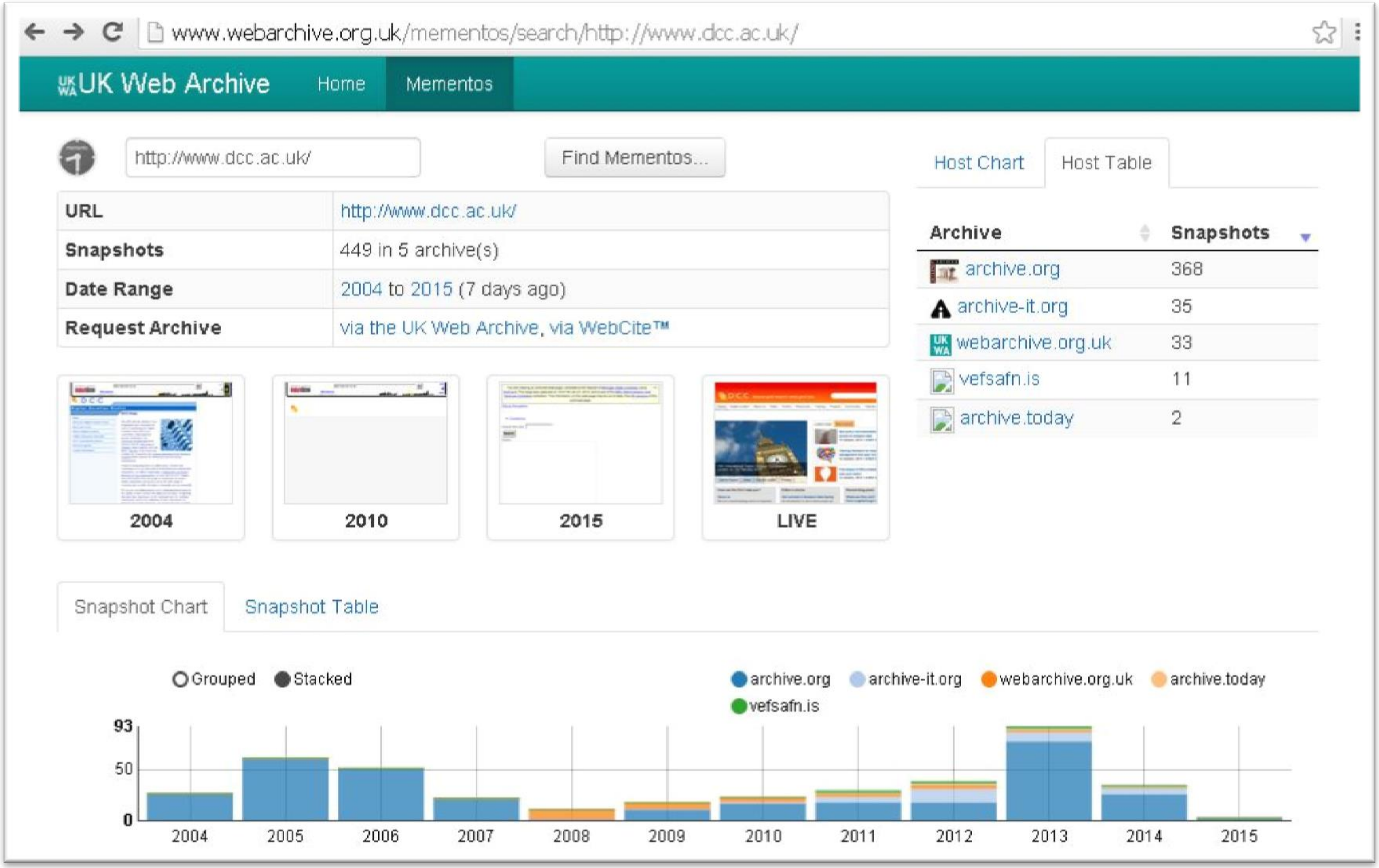
How it is done

How was the UK web linked in 1996?



- By Rainer Simon using UK Host-Level Link Graph (1996-2010) dataset.
- Based on the 1996 portion: 58,842 hosts (nodes); 184,433 host-to-host links (edges)
- UK web as part of the global web
- Scalability issues with large dataset over time

Mementos service (http://www.dcc.ac.uk)



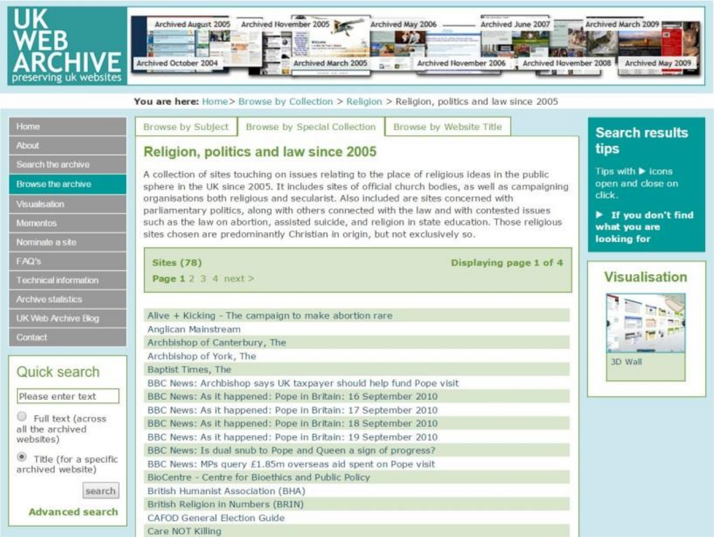
http://www.dcc.ac.uk/

URL	http://www.dcc.ac.uk/
Snapshots	449 in 5 archive(s)
Date Range	2004 to 2015 (7 days ago)
Request Archive	via the UK Web Archive, via WebCite™

Archive	Snapshots
archive.org	368
archive-it.org	35
webarchive.org.uk	33
vefsafn.is	11
archive.today	2

Collaboration with researchers

- Building collections
  - Researchers' involvement in scoping collections, selecting and describing websites
  - Creation of specific, (narrow) topical collections



You are here: Home> Browse by Collection > Religion > Religion, politics and law since 2005

Religion, politics and law since 2005

A collection of sites touching on issues relating to the place of religious ideas in the public sphere in the UK since 2005. It includes sites of official church bodies, as well as campaigning organisations both religious and secularist. Also included are sites concerned with parliamentary politics, along with others connected with the law and with contested issues such as the law on abortion, assisted suicide, and religion in state education. Those religious sites chosen are predominantly Christian in origin, but not exclusively so.

Sites (78) **Displaying page 1 of 4**

Page 1 2 3 4 next >

Alive + Kicking - The campaign to make abortion rare
Anglican Mainstream
Archbishop of Canterbury, The
Archbishop of York, The
Baptist Times, The
BBC News: Archbishop says UK taxpayer should help fund Pope visit
BBC News: As it happened: Pope in Britain: 16 September 2010
BBC News: As it happened: Pope in Britain: 17 September 2010
BBC News: As it happened: Pope in Britain: 18 September 2010
BBC News: As it happened: Pope in Britain: 19 September 2010
BBC News: Is dual snub to Pope and Queen a sign of progress?
BBC News: MPs query £1.85m overseas aid spent on Pope visit
BioCentre - Centre for Bioethics and Public Policy
British Humanist Association (BHA)
British Religion in Numbers (BRIN)
CAFOD General Election Guide
Care NOT Killing

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

Search results tips

Tips with ▶ icons open and close on click.

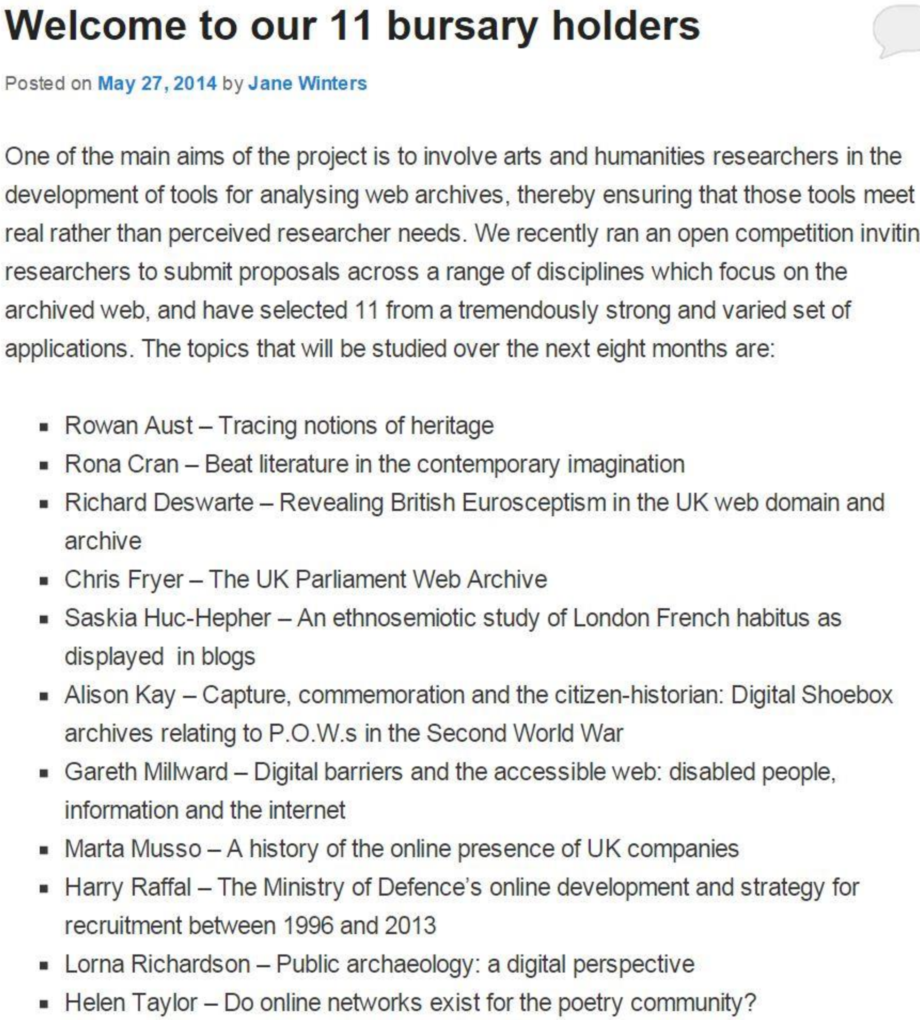
▶ **If you don't find what you are looking for**

- Formulating research question
  - Brain-storm sessions, workshops, discussion, surveys etc.
  - Lack of awareness & baseline knowledge
  - Challenging: you don't know what you don't know
  - Analytical access: web archive as dataset

Co-development

- Institute of Historical Research project: Analytical Access to the Domain Dark Archive (AADDA)
  - Experimental user interface
- Big UK Domain Data for Arts and Humanities (2014-2015)
  - Institution of Historical Research, Oxford Internet Institute, British Library and Aarhus University
  - Develop theoretical and methodological framework for the study of web archives
  - Build on ADDAA: researchers and the BL co-produce access tools
  - Study of the history of UK web space from 1996 to 2013
  - Research use cases
- Changing the way we archive the web

Web archiving researcher bursaries



Welcome to our 11 bursary holders

Posted on May 27, 2014 by Jane Winters

One of the main aims of the project is to involve arts and humanities researchers in the development of tools for analysing web archives, thereby ensuring that those tools meet real rather than perceived researcher needs. We recently ran an open competition inviting researchers to submit proposals across a range of disciplines which focus on the archived web, and have selected 11 from a tremendously strong and varied set of applications. The topics that will be studied over the next eight months are:

- Rowan Aust – Tracing notions of heritage
- Rona Cran – Beat literature in the contemporary imagination
- Richard Deswarte – Revealing British Eurosceptism in the UK web domain and archive
- Chris Fryer – The UK Parliament Web Archive
- Saskia Huc-Hepher – An ethnosemiotic study of London French habitus as displayed in blogs
- Alison Kay – Capture, commemoration and the citizen-historian: Digital Shoebox archives relating to P.O.W.s in the Second World War
- Gareth Millward – Digital barriers and the accessible web: disabled people, information and the internet
- Marta Musso – A history of the online presence of UK companies
- Harry Raffal – The Ministry of Defence's online development and strategy for recruitment between 1996 and 2013
- Lorna Richardson – Public archaeology: a digital perspective
- Helen Taylor – Do online networks exist for the poetry community?

INSTITUTE OF HISTORICAL RESEARCH | University of London School of Advanced Study

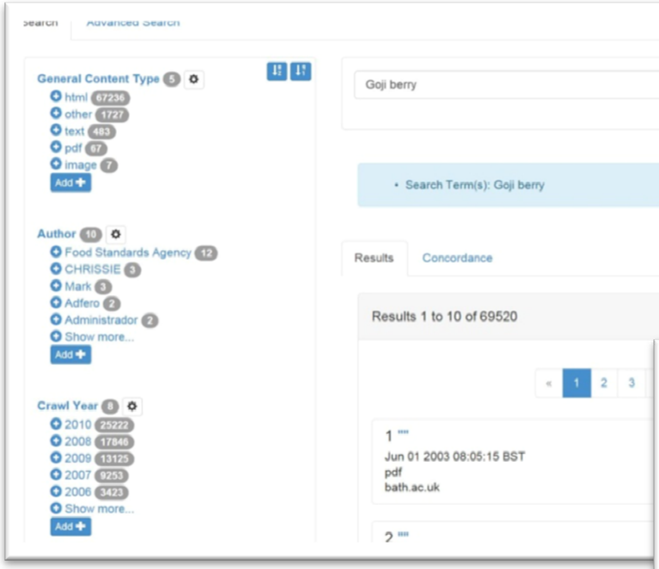
BRITISH LIBRARY

oxford internet institute university of oxford

AARHUS UNIVERSITY

Arts & Humanities Research Council

Shine

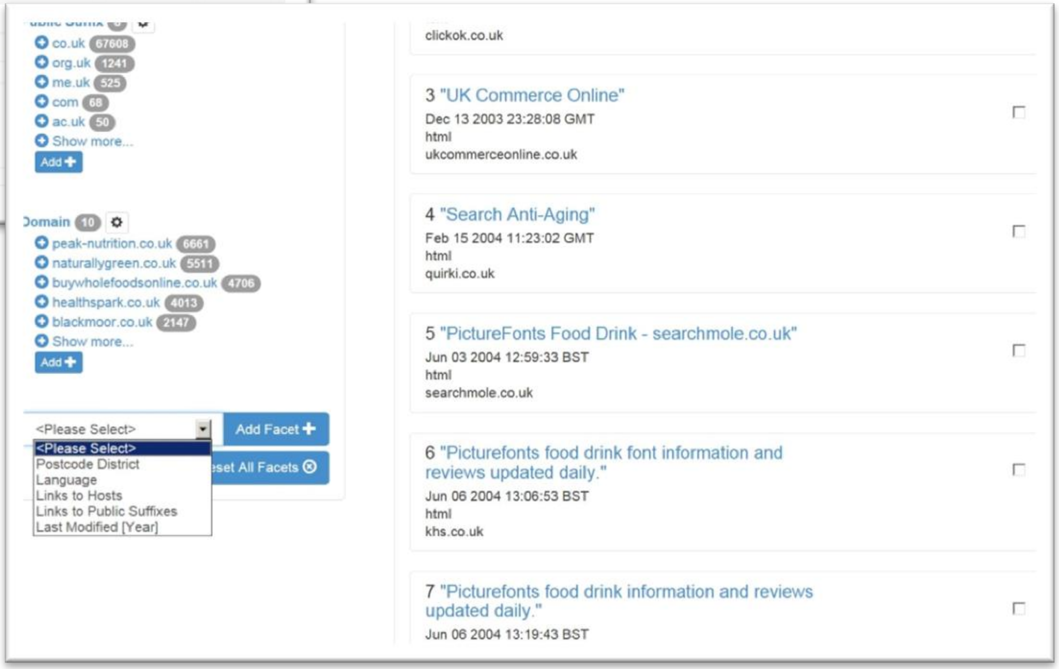


- Search Term(s): Goji berry

Results 1 to 10 of 69520

1 ""
Jun 01 2003 08:05:15 BST
pdf
bath.ac.uk

2 ""



clickok.co.uk

3 "UK Commerce Online" □
Dec 13 2003 23:28:08 GMT
html
ukcommerceonline.co.uk

4 "Search Anti-Aging" □
Feb 15 2004 11:23:02 GMT
html
quirki.co.uk

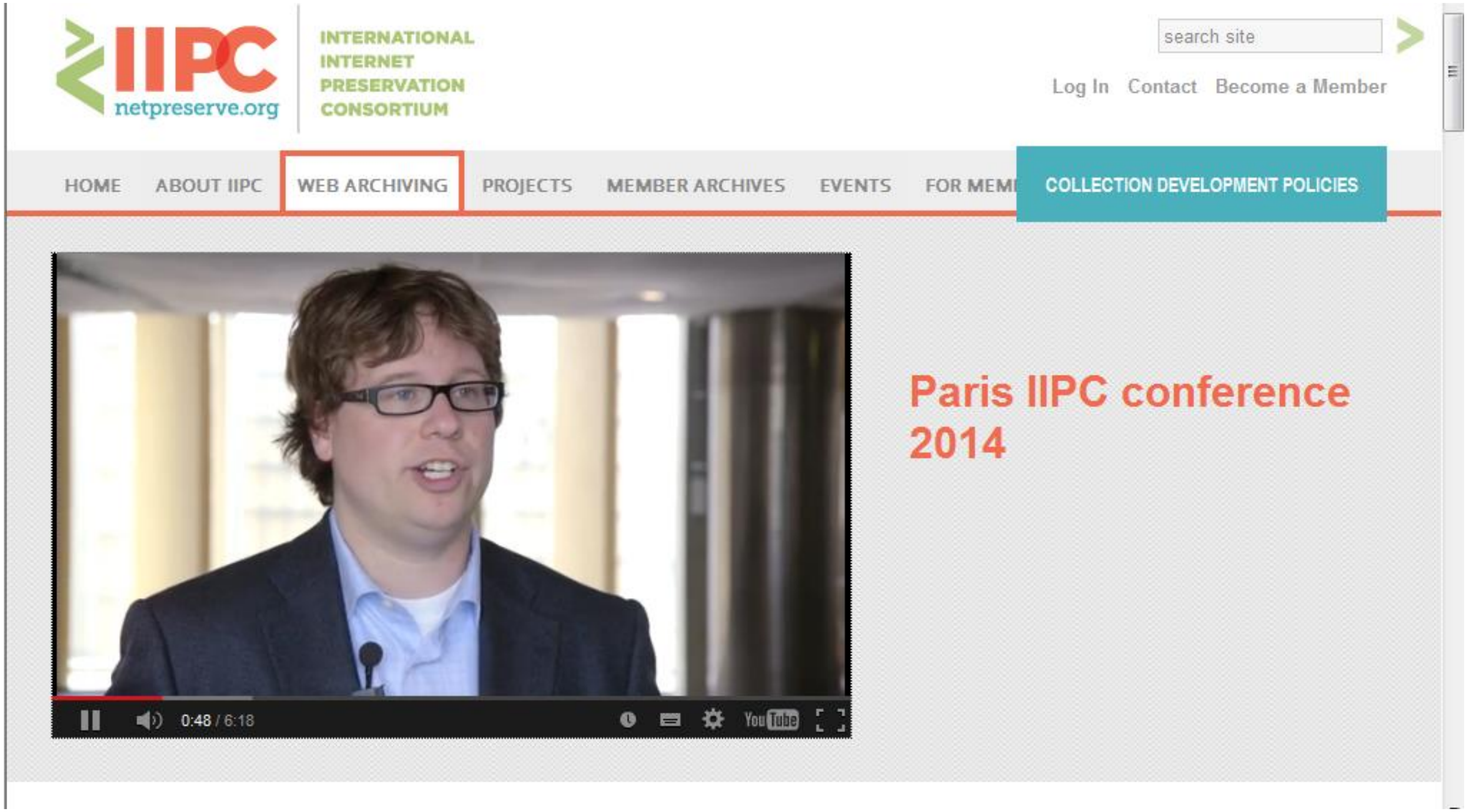
5 "PictureFonts Food Drink - searchmole.co.uk" □
Jun 03 2004 12:59:33 BST
html
searchmole.co.uk

6 "Picturefonts food drink font information and reviews updated daily." □
Jun 06 2004 13:06:53 BST
html
khs.co.uk

7 "Picturefonts food drink information and reviews updated daily." □
Jun 06 2004 13:19:43 BST

- Query building
- Corpus formation and handling
- Annotation and curation
- In-corpus analysis
- Whole-dataset analysis

Scholarly use of web archives



http://netpreserve.org/web-archiving/videos

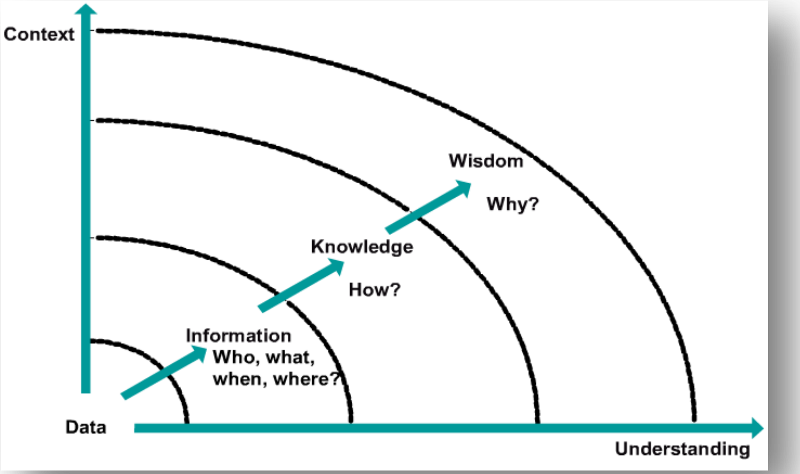
Requirements for web archives

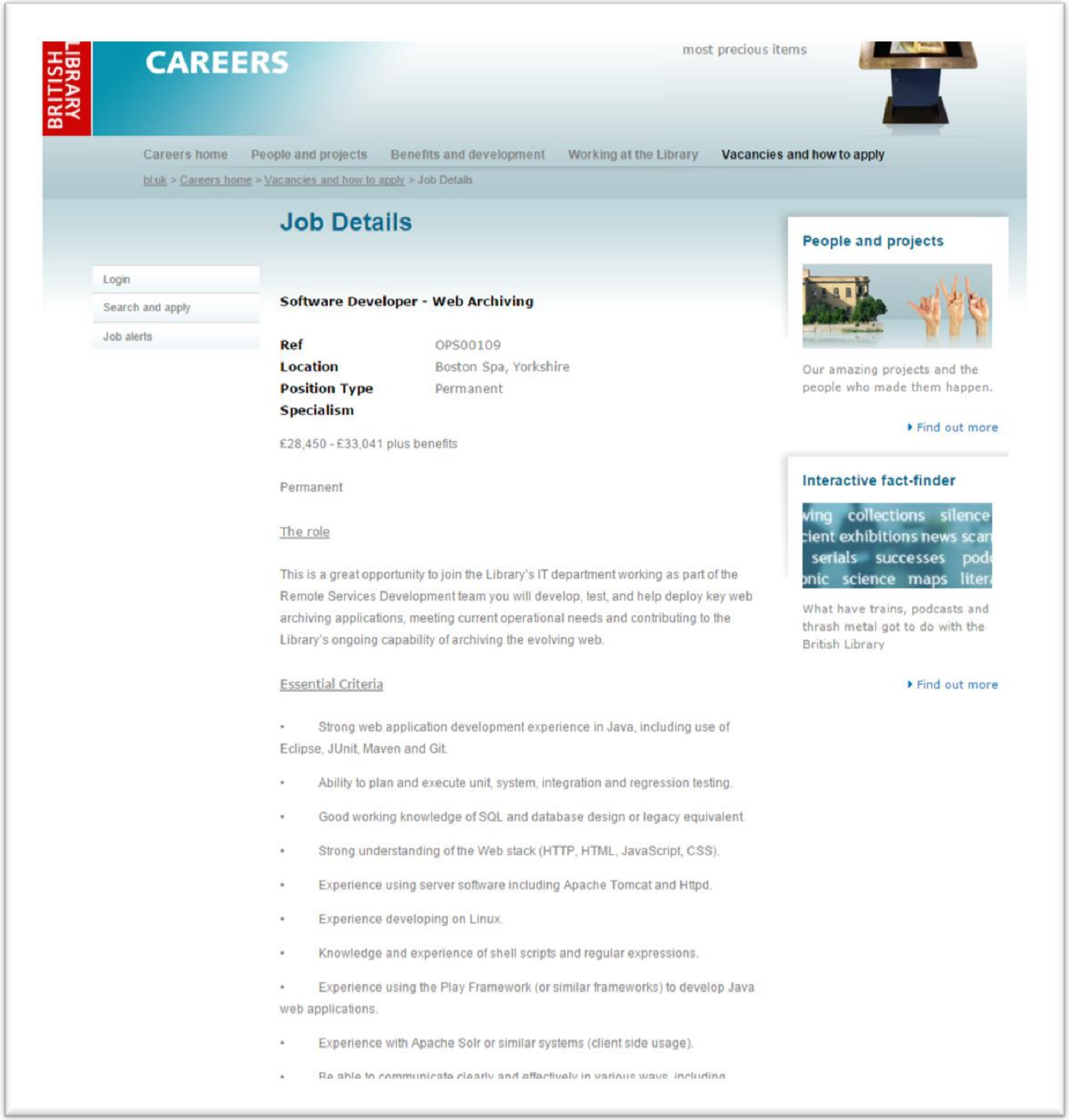
Scholarly requirements	Requirements for web archives
Availability	No access restriction, available online
Text, paratext and context	- Access to text and any contextual information at various granularity. - Access to secondary datasets
Persistence and citability	- Persistent identifiers - Standards of citing archived websites - Integration with bibliographical management tools (eg Zotero)
Formation of research corpus	- Archiving of research corpora on demand - Means to mix, match and reassemble corpora
Quality	- Represents as much as possible the live website in completeness, intellectual content, behaviour and look and feel - Information on what is “missing” to help interpret incomplete sources
Close and distant reading	Multiple access methods including tools for data analytics and visualisations
Boundary & format-independent	- Interlinked web archives - integration with other digital and printed sources

IDCC 2025

- The UK Web Archive
  - Single search for all collections
  - Base-line knowledge self-explanatory, text and much contextual information
  - Link to many more web archives
  - Inactive websites highlighted
  - Open datasets / content

- "Macroscope" of the UK web history
  - "a single data point, .. both visualised at scale in the context of a billion other data points, and drilled down to its smallest compass"
- Flexibility to reconstruct context, big or small
- Seamlessly linked web archives
- Crawlers as capable as browsers
- Independent use of web archives




BRITISH LIBRARY
CAREERS
most precious items
Careers home
People and projects
Benefits and development
Working at the Library
Vacancies and how to apply
bl.uk > Careers home > Vacancies and how to apply > Job Details
Login
Search and apply
Job alerts
People and projects
Our amazing projects and the people who made them happen.
Find out more
Interactive fact-finder
collections silence exhibitions news serials successes science maps
What have trains, podcasts and thrash metal got to do with the British Library
Find out more

Job Details

Software Developer - Web Archiving

Ref OPS00109
Location Boston Spa, Yorkshire
Position Type Permanent
Specialism

£28,450 - £33,041 plus benefits

Permanent

The role

This is a great opportunity to join the Library's IT department working as part of the Remote Services Development team you will develop, test, and help deploy key web archiving applications, meeting current operational needs and contributing to the Library's ongoing capability of archiving the evolving web.

Essential Criteria

- Strong web application development experience in Java, including use of Eclipse, JUnit, Maven and Git.
- Ability to plan and execute unit, system, integration and regression testing.
- Good working knowledge of SQL and database design or legacy equivalent.
- Strong understanding of the Web stack (HTTP, HTML, JavaScript, CSS).
- Experience using server software including Apache Tomcat and Httpd.
- Experience developing on Linux.
- Knowledge and experience of shell scripts and regular expressions.
- Experience using the Play Framework (or similar frameworks) to develop Java web applications.
- Experience with Apache Solr or similar systems (client side usage).
- Be able to communicate clearly and effectively in various ways, including

Java Developer for Web Archiving

Closing date Midnight 15th February

Please help spread the word or Apply!

Contact: Helen.hockx-yu@bl.uk
Twitter: @ukwebarchive

References

- Ainsworth, Scott G., AlSum, Ahmed, SalahEldeen, Hany, Weigle, Michele C., Nelson, Michael L. (2012) ‘How much of the web is archived?’, http://arxiv.org/abs/1212.6177 .

- Brügger, Niels (2010) *Website Analysis: Elements of a conceptual architecture*. Aarhus: The Centre for Internet Research.

- Gomes, D., Miranda, J., Costa, M., A survey on web archiving initiatives. In *Proc.* TPDL'11, Springer-Verlag Berlin, Heidelberg (2011), 408-420, http://dl.acm.org/citation.cfm?id=2042590 .

- Jacobson, P (2014), ‘Inside the Struggle to Preserve the World's Data’. http://www.newsweek.com/2014/07/11/inside-struggle-preserve-worlds-data-257020.html.

- Moretti, Franco (2000) ‘Conjectures on world literature’. *New Left Review*, 1, http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature.