Collecting and Providing Access to Large Scale Archived Web Data

Helen Hockx-Yu

Head of Web Archiving, British Library

Web Archives – key characteristics

- Snapshots of web resources, taken at given point in time – a living thing
- Archived web resources are “reborn” – different from digitised and born digital collections (Brugger, N., 2013)
- Different from the live web in many ways
- Additional temporal aspects: many different versions of the same thing? Different things in own right?
- Have boundaries & limitations, determined by purpose, strategy and technological choices

Different 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 Archiving at the British Library

- Started web archiving in 2003: Open UK Web Archive
  - Selective, topical collections and key sites
  - 20TB, 14,792 websites
- Archiving UK Web for non-print Legal Deposit since April 2013: Legal Deposit UK Web Archive
  - Comprehensive national archive with on-site access only
  - 2013 UK domain crawl contains 31TB, 1.6 billion URLs from 3.8 million domains
- Collect UK digital heritage and provide continued access to archived web resources

UK websites – territoriality explained

An online work is considered as “published in the UK” and therefore in scope for Legal Deposit, if it meets either of the following criteria:

(a) it is made available to the public from a website with a domain name which relates to the United Kingdom or to a place within the United Kingdom; or

(b) it is made available to the public by a person and any of that person’s activities relating to the creation or the publication of the work take place within the United Kingdom

The Legal Deposit Libraries (Non-Print Works) Regulations, 2013

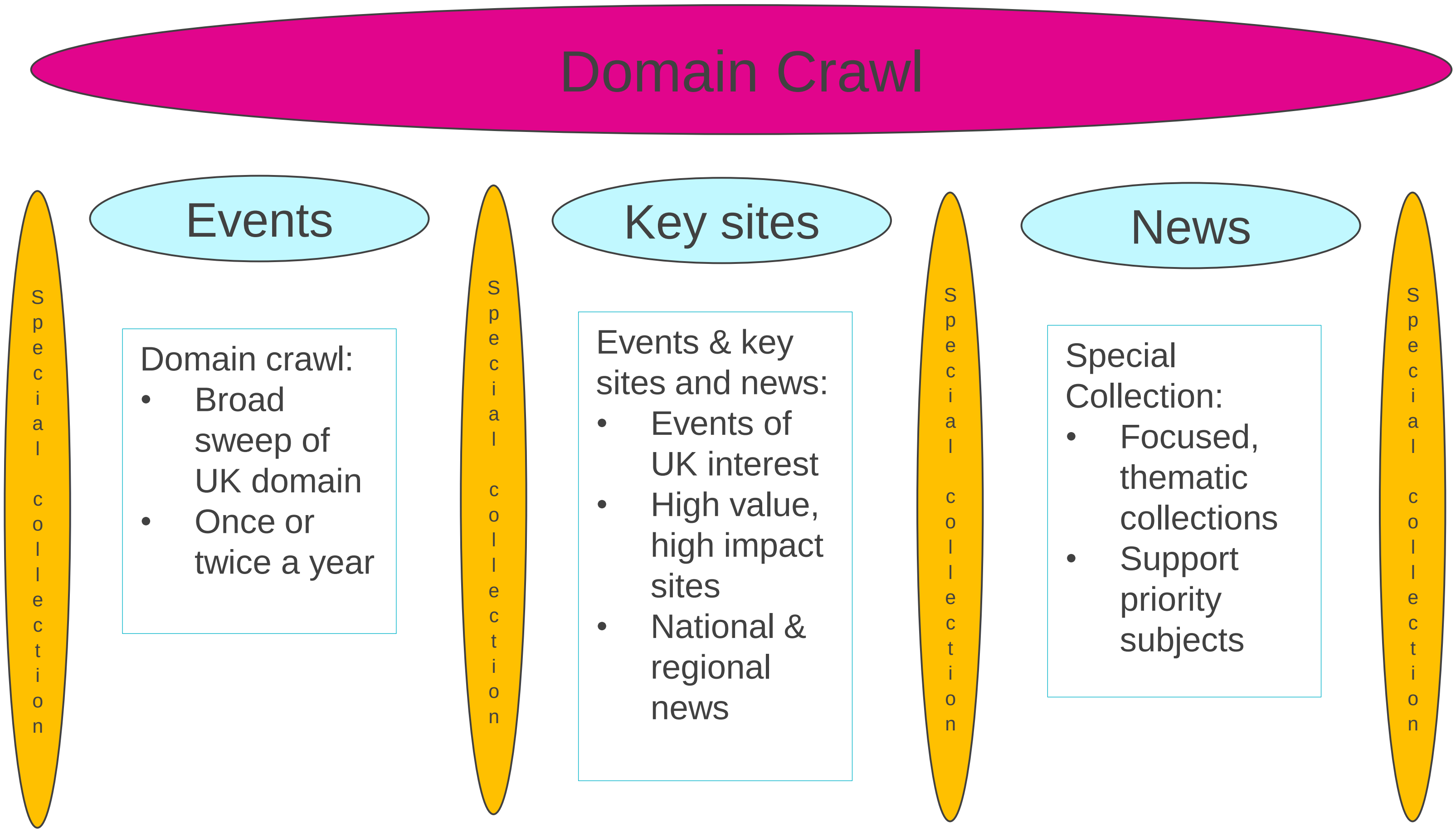
Territoriality - implementation

- All websites with a .uk domain name
- non .uk websites have to meet at least one criteria
  - UK Hosting: check external IP geo-location database and add in-scope URLs to the fetch-chain
  - UK postal address
  - Correspondence
  - Professional judgement

Legal requirements, policies and decisions

- Content not in scope
  - Sound recordings and films (unless incidental to the deposited work)
  - Intranet, personal emails
- Follow Robots Exclusion Protocol by default with some exceptions
  - Mutual agreement
  - All seed pages and embedded content (i.e., CSS, JavaScript and images)
- Host Cap
  - A 512MB ‘cap’ per host based on the known site size distribution
  - Tail of large sites based on curators’ assessment
  - Some excepted and crawled without data cap
- Collect embedded content regardless where it is hosted

Collecting strategy for websites



Domain Crawl

Special collection

Events

Domain crawl:

- Broad sweep 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:

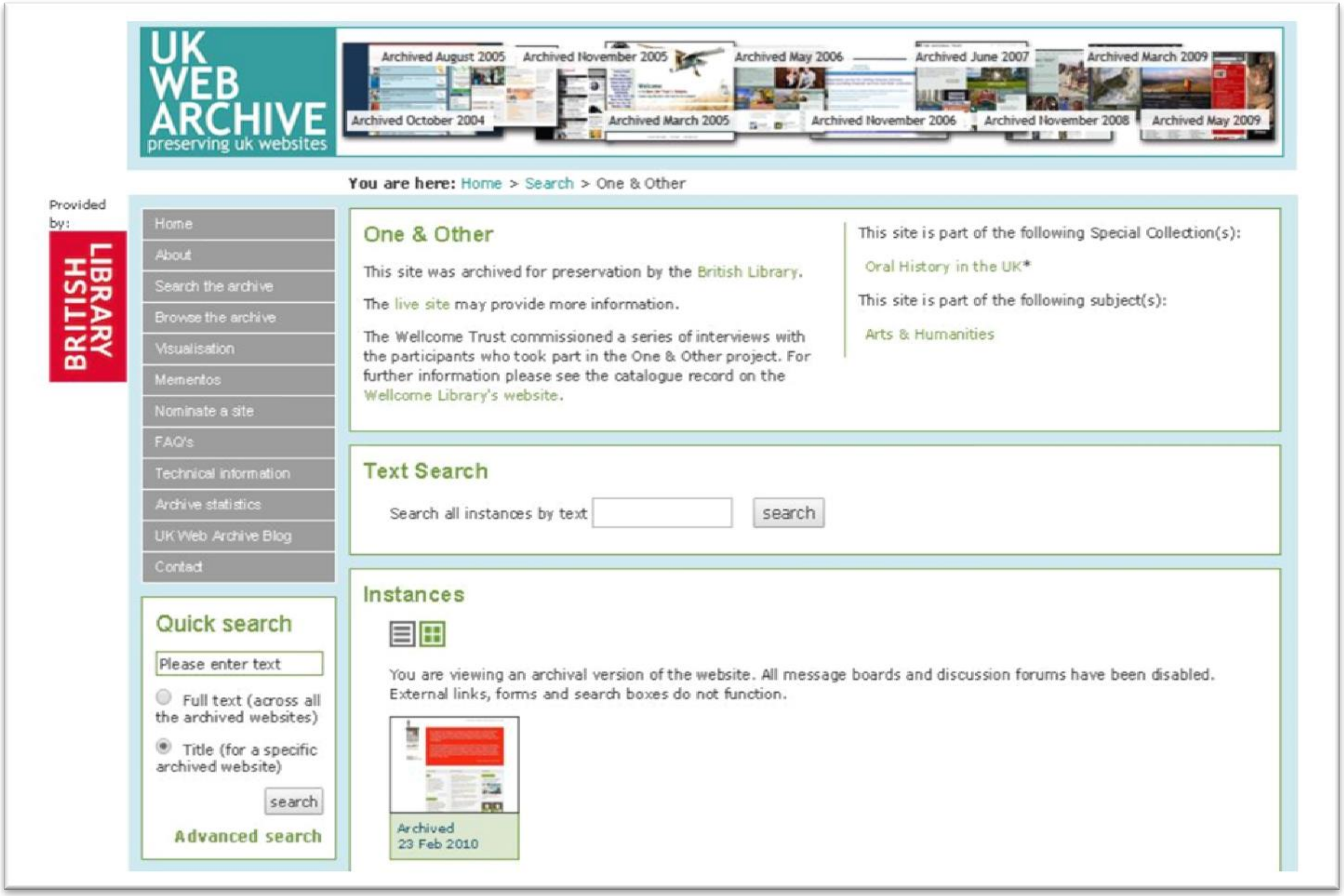
- Focused, thematic collections
- Support priority subjects

Special collection

Curating websites

- Concept of a “website” relates to the way people use and present information
- Collection is how libraries organise knowledge
- Description at website and collection levels including subject terms
- Presenting “collections” and “websites” to end-users

Website (landing page)



You are here: Home > Search > One & Other

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 ▭ search

Quick search

Please enter text

○ Full text (across all the archived websites)

● Title (for a specific archived website)

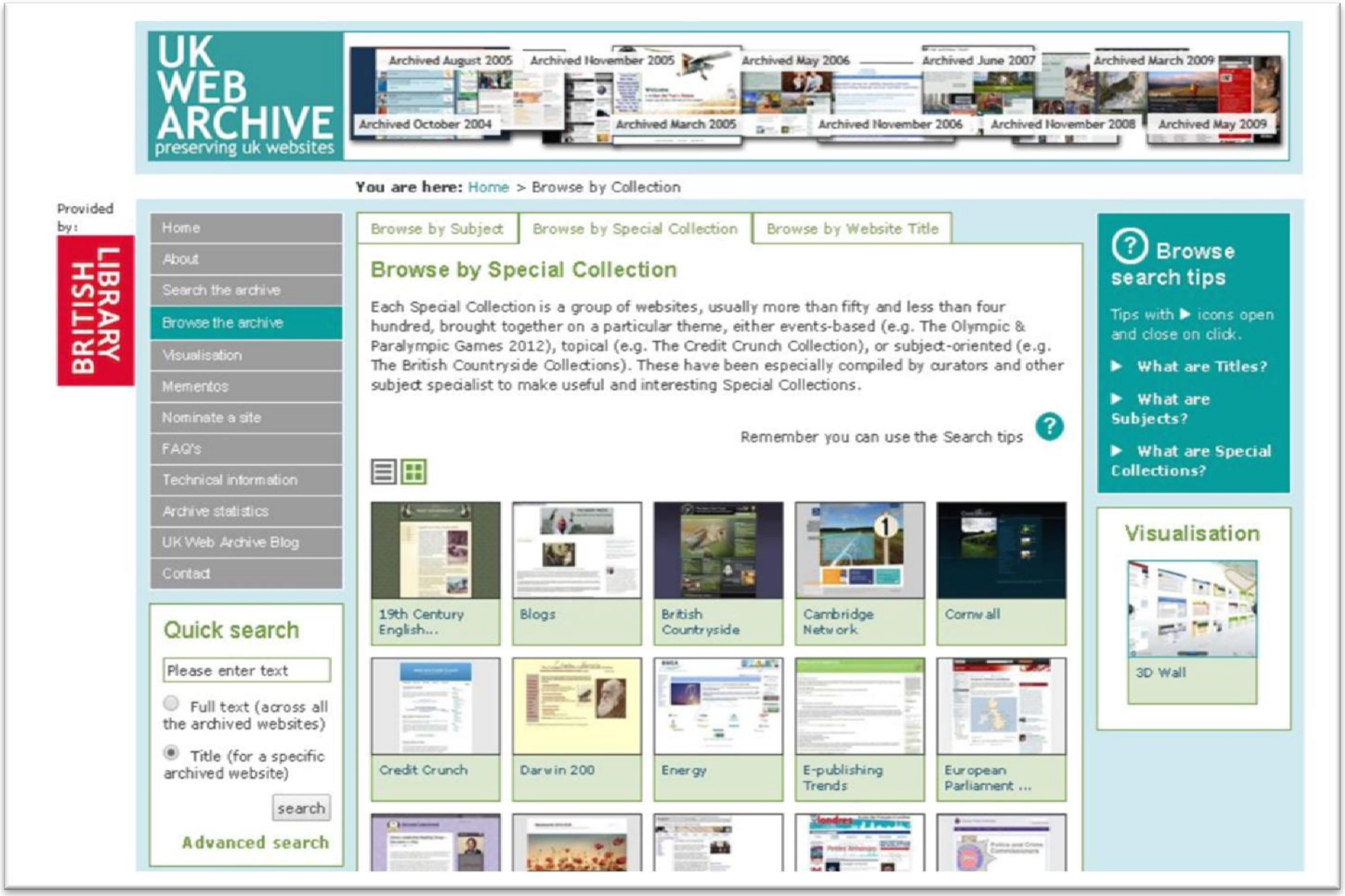
search

Advanced search

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.

Collections



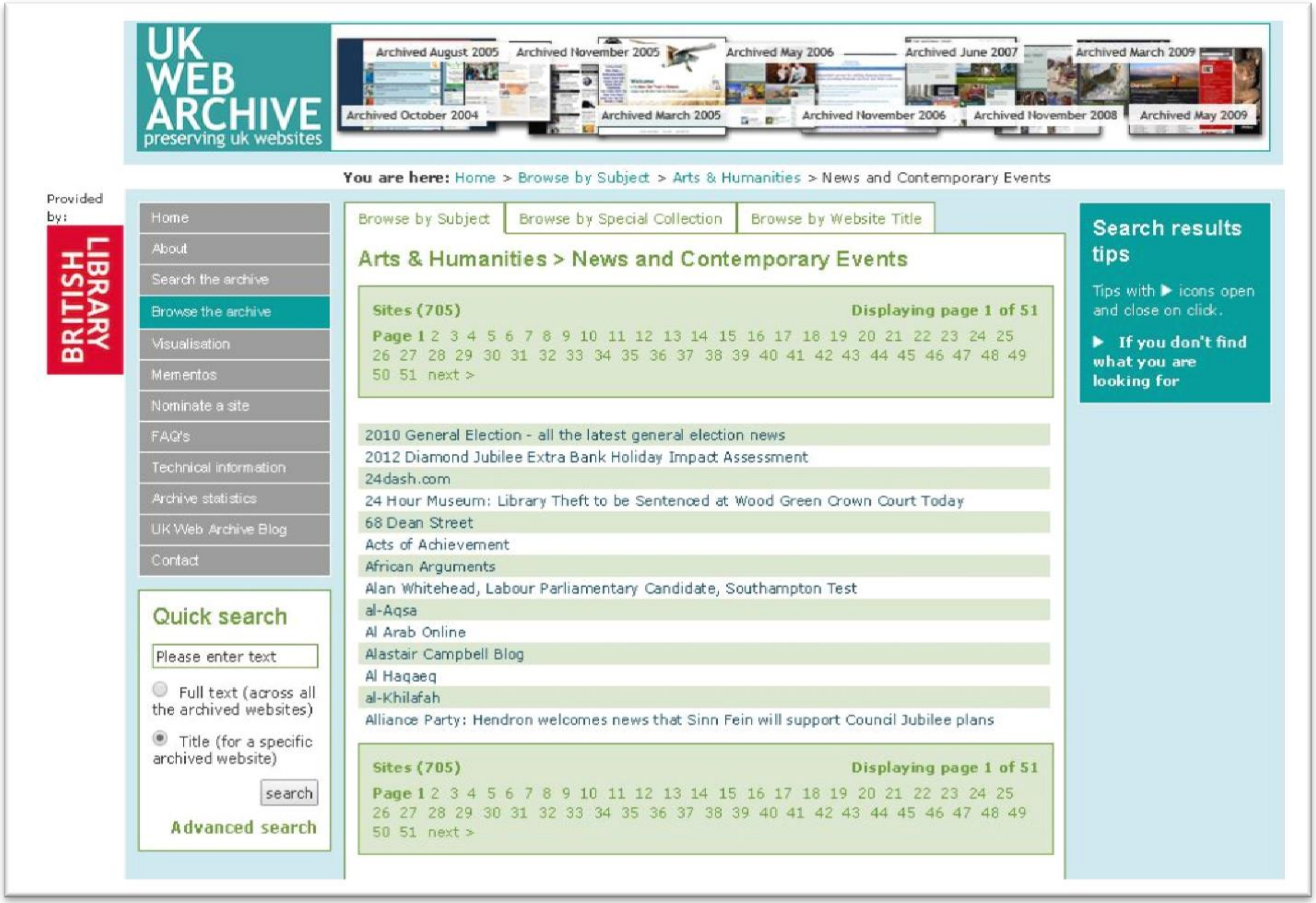
You are here: Home > Browse by Collection

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

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 >

○ 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

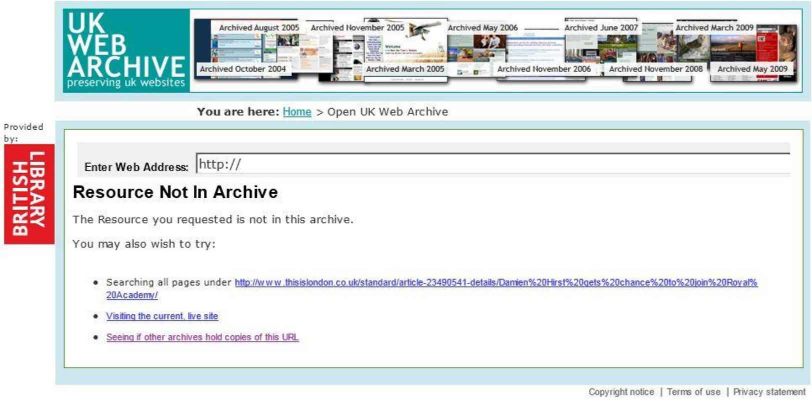
Full-text index and facets

- The UK Legal Deposit UK Web Archive offers full-text search. Results can be refined by
  - 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

“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

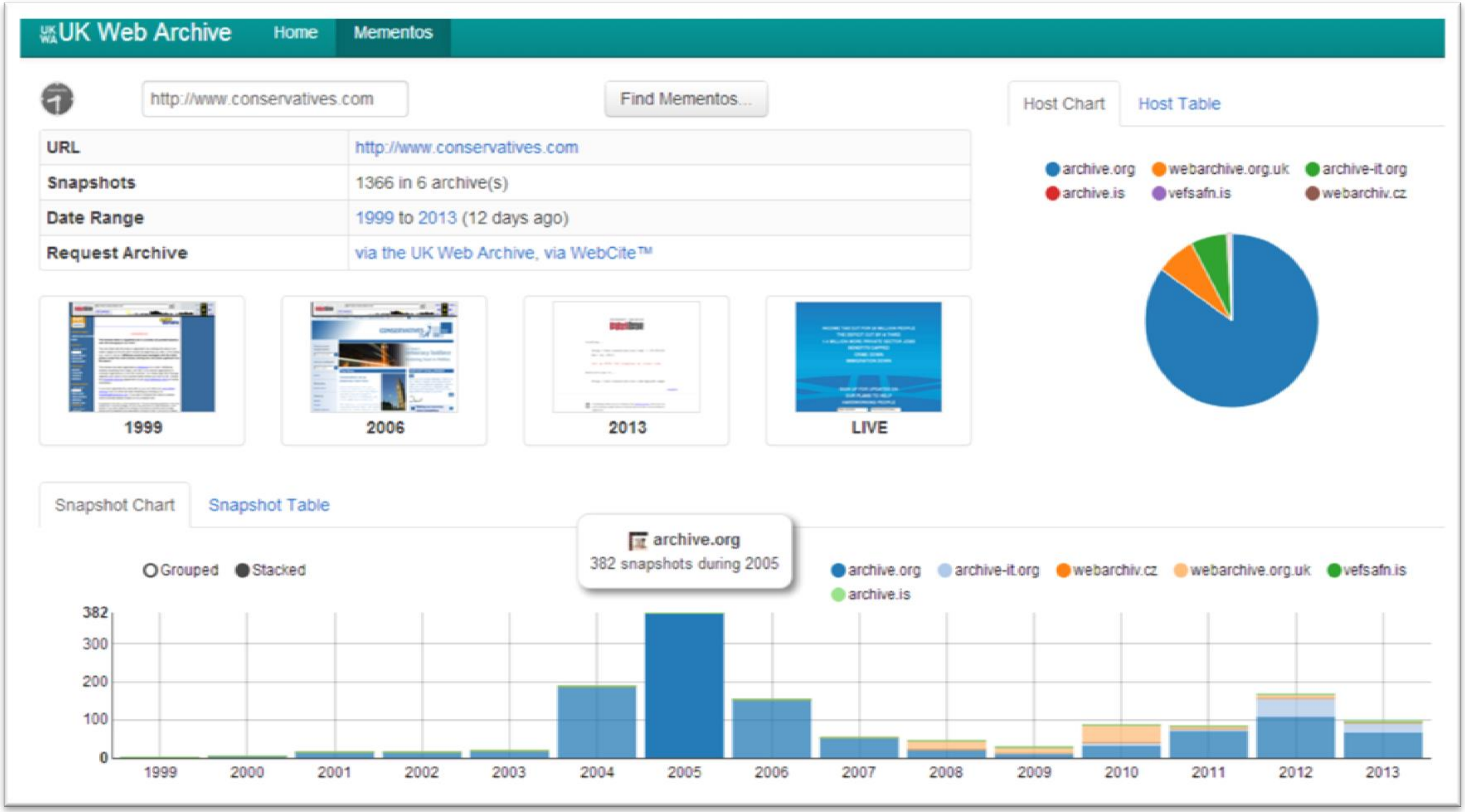
- Technical limitations, e.g. dynamic content the crawler could not collect

- Avoid dead end in navigation
  - – Search
  - – Link to live web
  - – Find archived copies elsewhere

Memento service

- Allow users to finding archived web pages (*mementos*) in multiple web archives across the world (search based on aggregated metadata)
  - Exposes the *memento protocol*, which adds time dimension to HTTP - accessing the past web as it is to access the current web
  - uses the Memento aggregate TimeGate hosted by lanl.gov
  - Source code
- Find memento bookmarklet, finding archived versions of 404 webpages while browsing

Who has archived http://www.conservatives.com/?



UK Web Archive | Home | Mementos

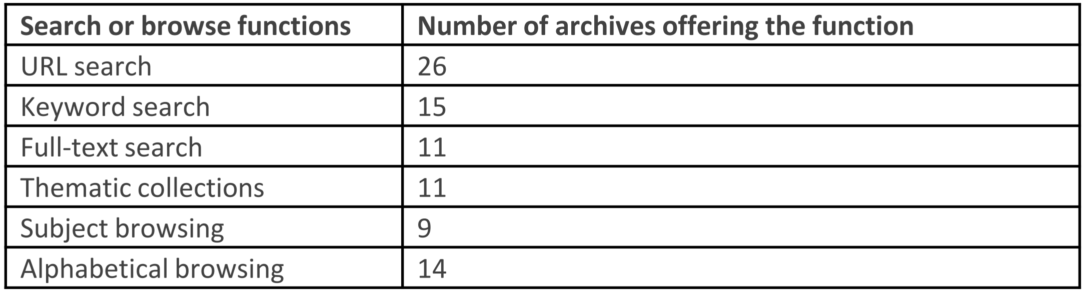
http://www.conservatives.com | Find Mementos...

URL	http://www.conservatives.com
Snapshots	1366 in 6 archive(s)
Date Range	1999 to 2013 (12 days ago)
Request Archive	via the UK Web Archive, via WebCite™

1999
2006
2013
LIVE
Host Chart
Host Table
● archive.org ● webarchive.org.uk ● archive-it.org
● archive.is ● vefsafn.is ● webarchiv.cz
Snapshot Chart
Snapshot Table
○ Grouped ● Stacked
archive.org
382 snapshots during 2005
● archive.org ● archive-it.org ● webarchiv.cz ● webarchive.org.uk ● vefsafn.is
● archive.is
382
300
200
100
0
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Access and use of web archives

- Problematic at two levels: legal restriction and (single) envisaged use case
- Archive webpages as historical documents used for reference
- URL search is the standard, universal access method



Search or browse functions	Number of archives offering the function
URL search	26
Keyword search	15
Full-text search	11
Thematic collections	11
Subject browsing	9
Alphabetical browsing	14

Access methods of 29 IIPC members' web archives

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.
- “Distant reading” (Franco Moretti) focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems

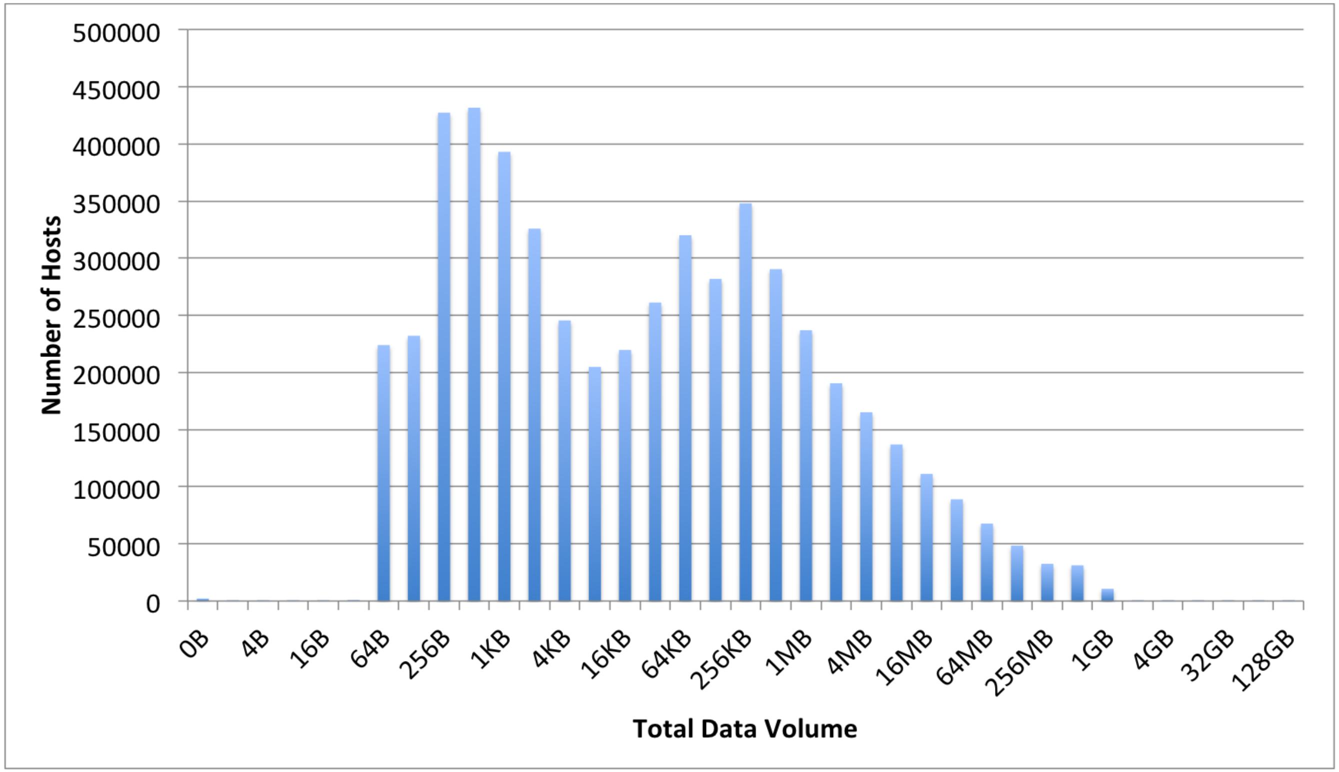
There is so much more…

- Statistical overview, scale and distribution of a web national domain
- Size: bytes
- Space: geo location, postcodes
- Type of content, eg file format, language
- Structure, linked entities and networks
- Evolution, pattern of change over time, eg domain names
- Correlation, eg between certain term and historical event
- Also viral content, crawl logs with eg http response codes

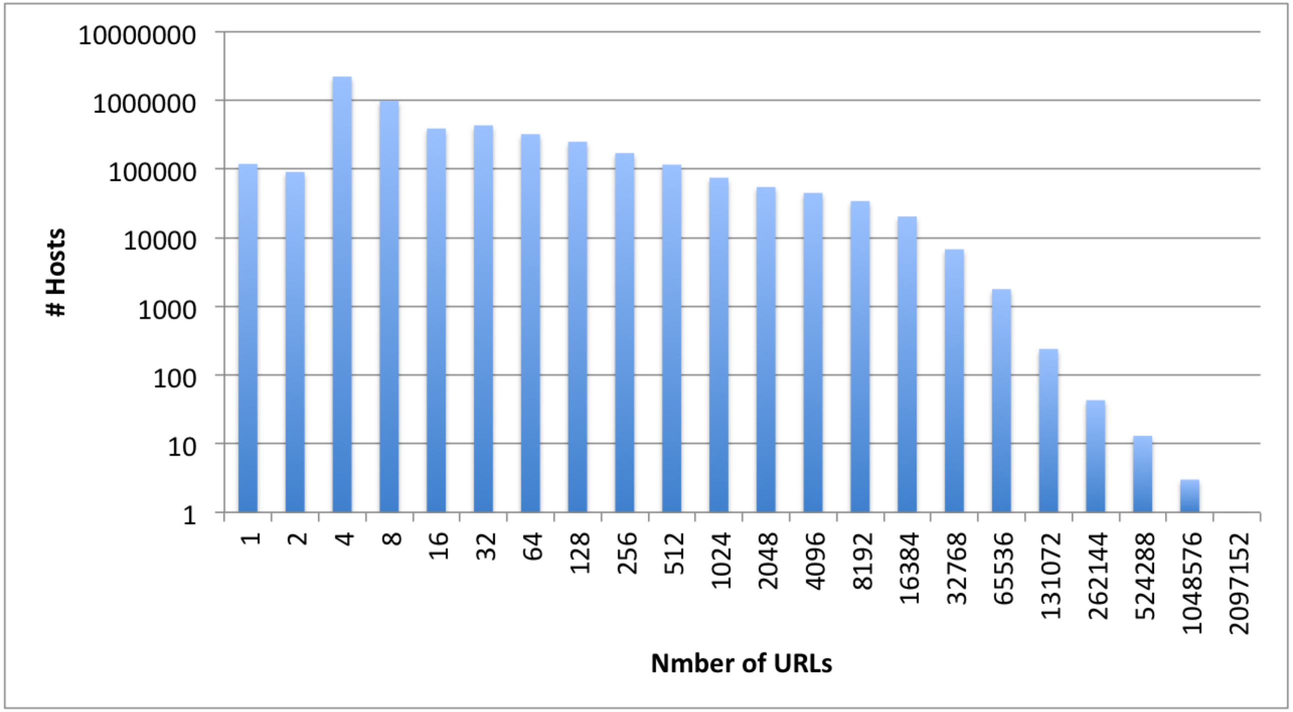
JISC UK web domain dataset (1996-2013)

- Collaboration between the Internet Archive (IA), the Joint Information Systems Committee (JISC) and the British Library
- Extracted copies of UK websites from the Internet Archives collection
  - 1st tranche : 1996 – 2010, 30TB, 2.5 billion URLs
  - 2nd tranche: 2010 – April 2013, 27.5TB, 1.5 billion URLs (estimated)
- Research agreement between JISC and IA, upholding IA's Terms of Use
  - Access via IA's Wayback Machine
  - Allows replication / extraction of derivative or secondary datasets , some available at http://data.webarchive.org.uk/opendata/ukwa.ds.2/
- Test-bed to develop new capability and services

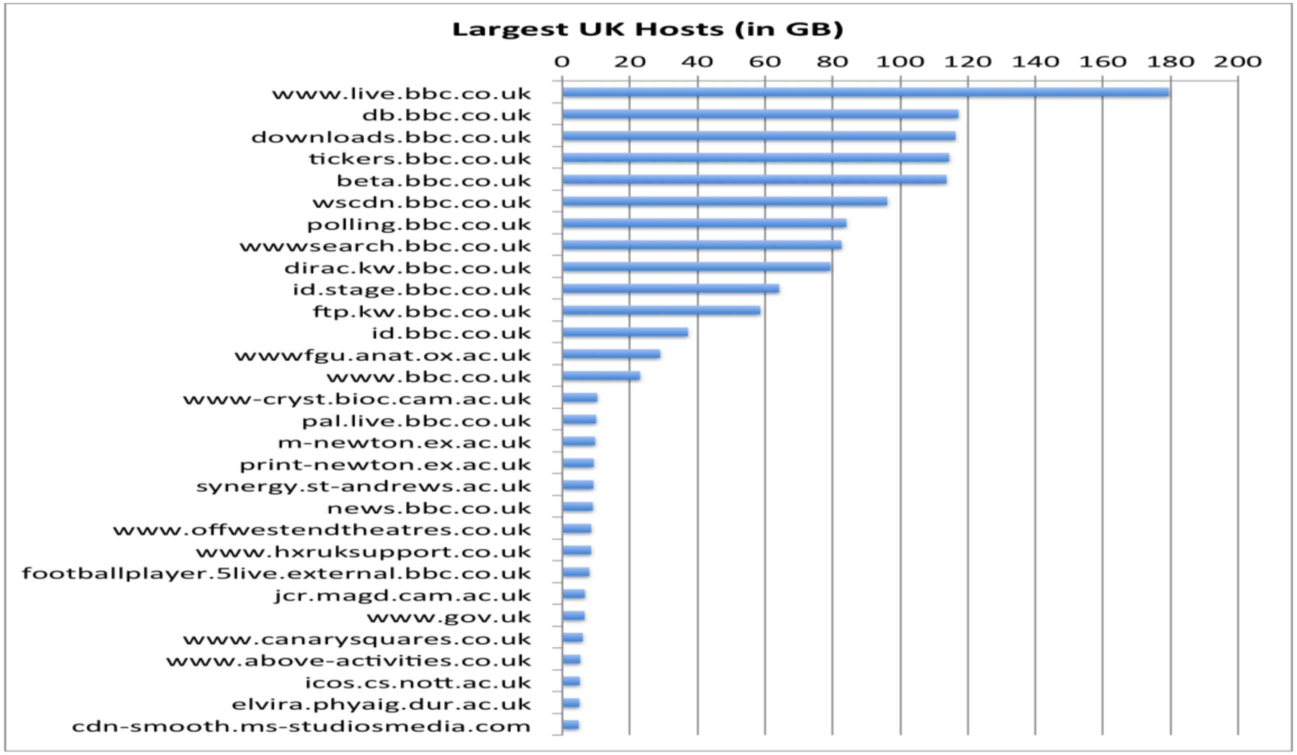
UK Web 2013: data volume per host



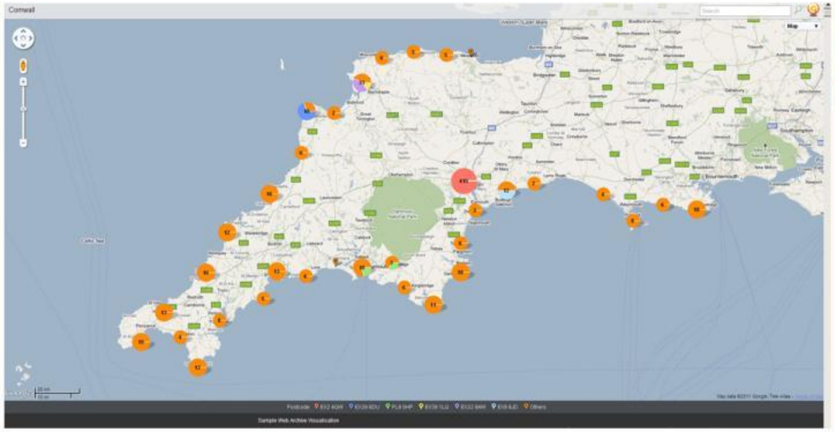
UK Web 2013: number of URLs per host

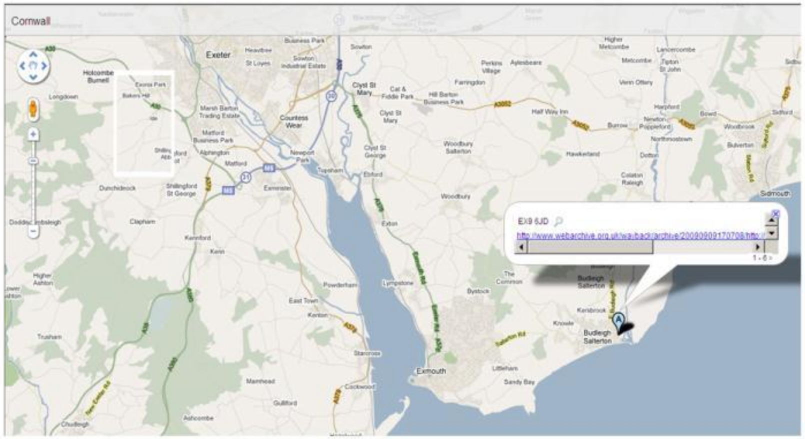


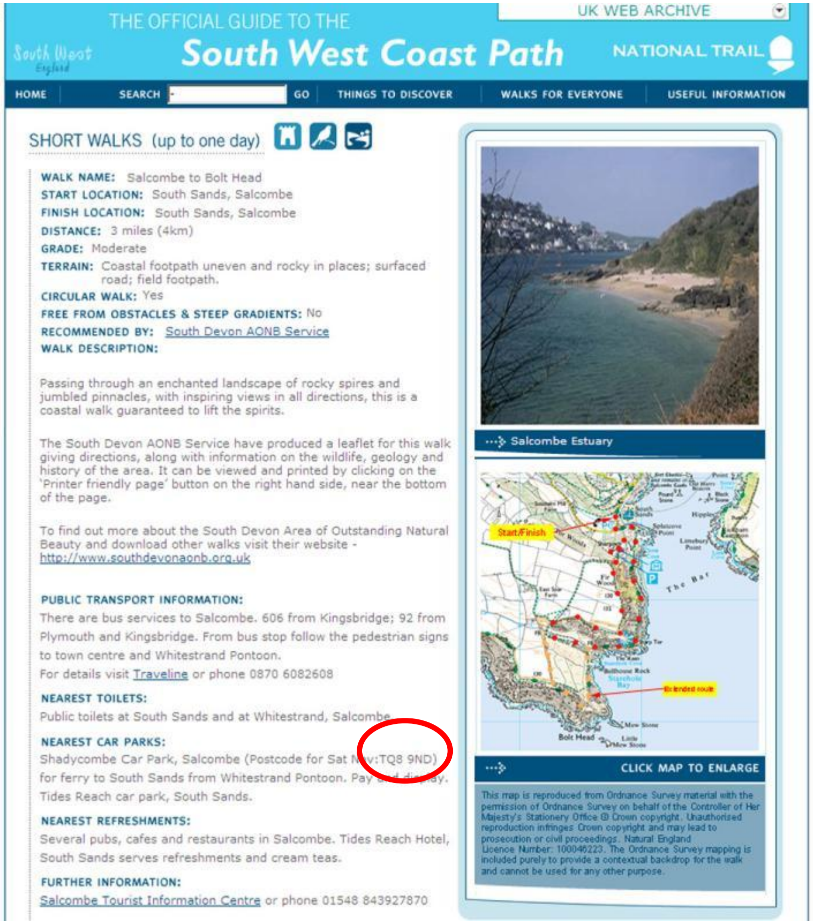
UK Web 2013: the largest host



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

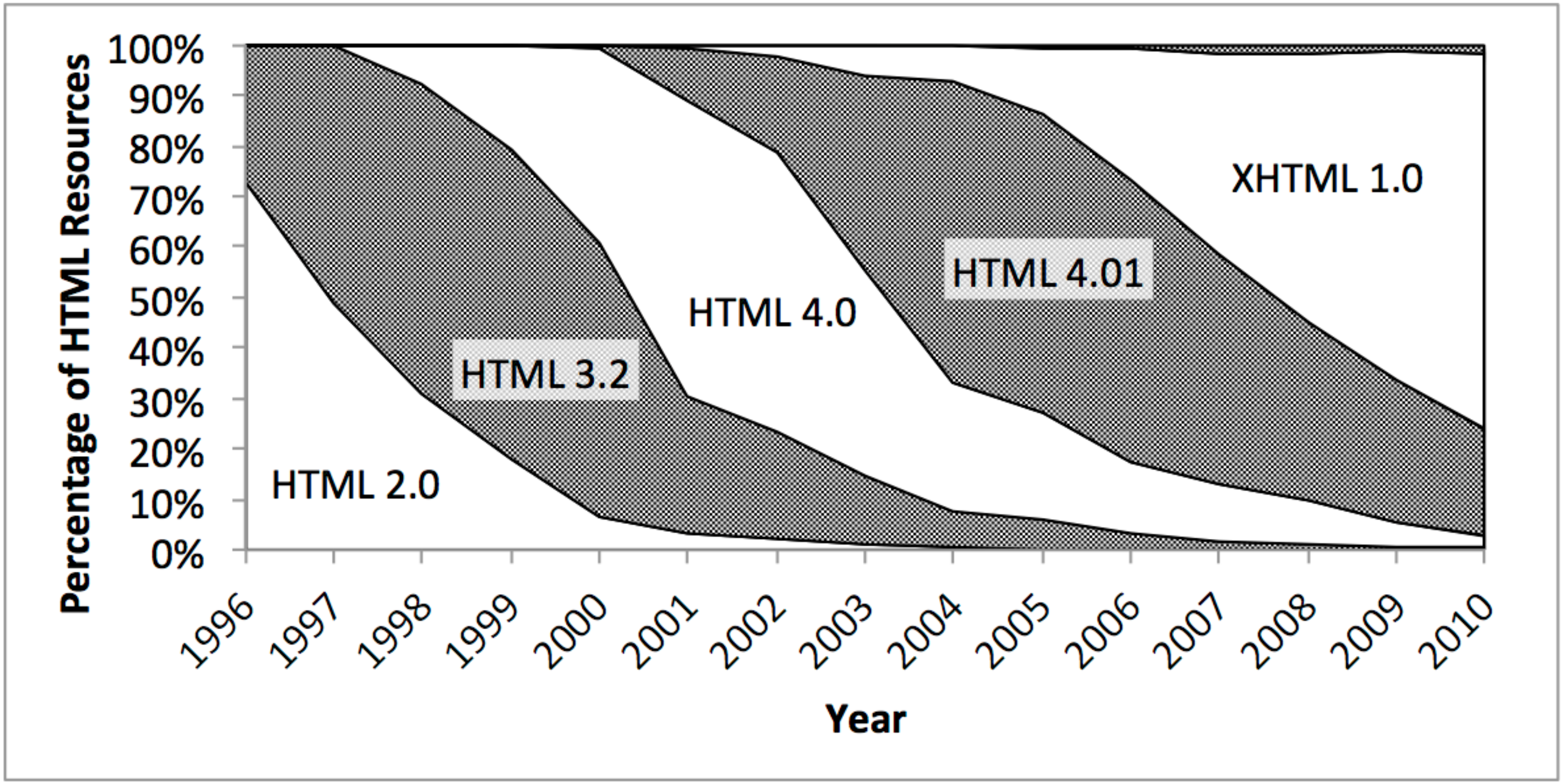
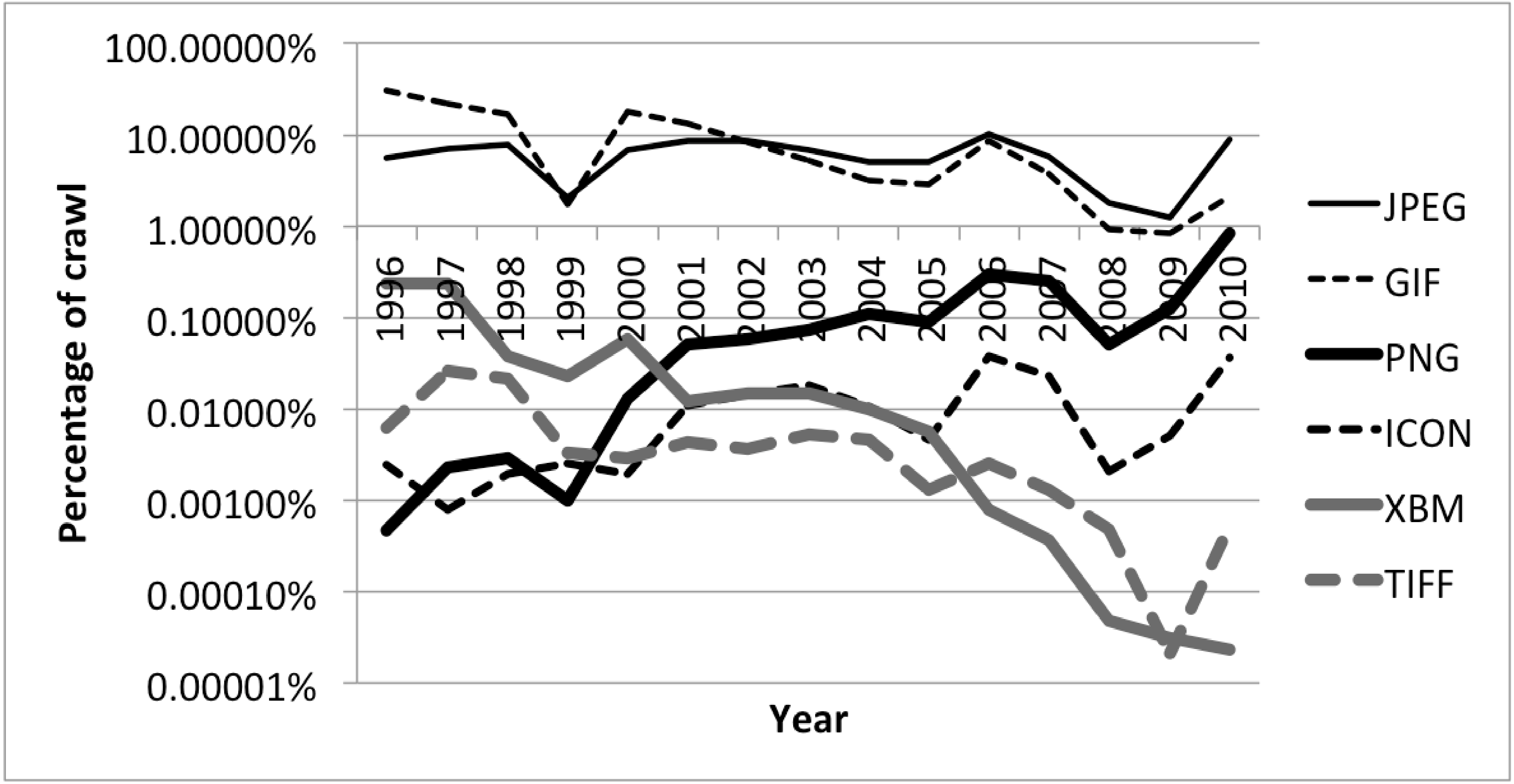


Image Format Analysis



Exploring Host Link Graph

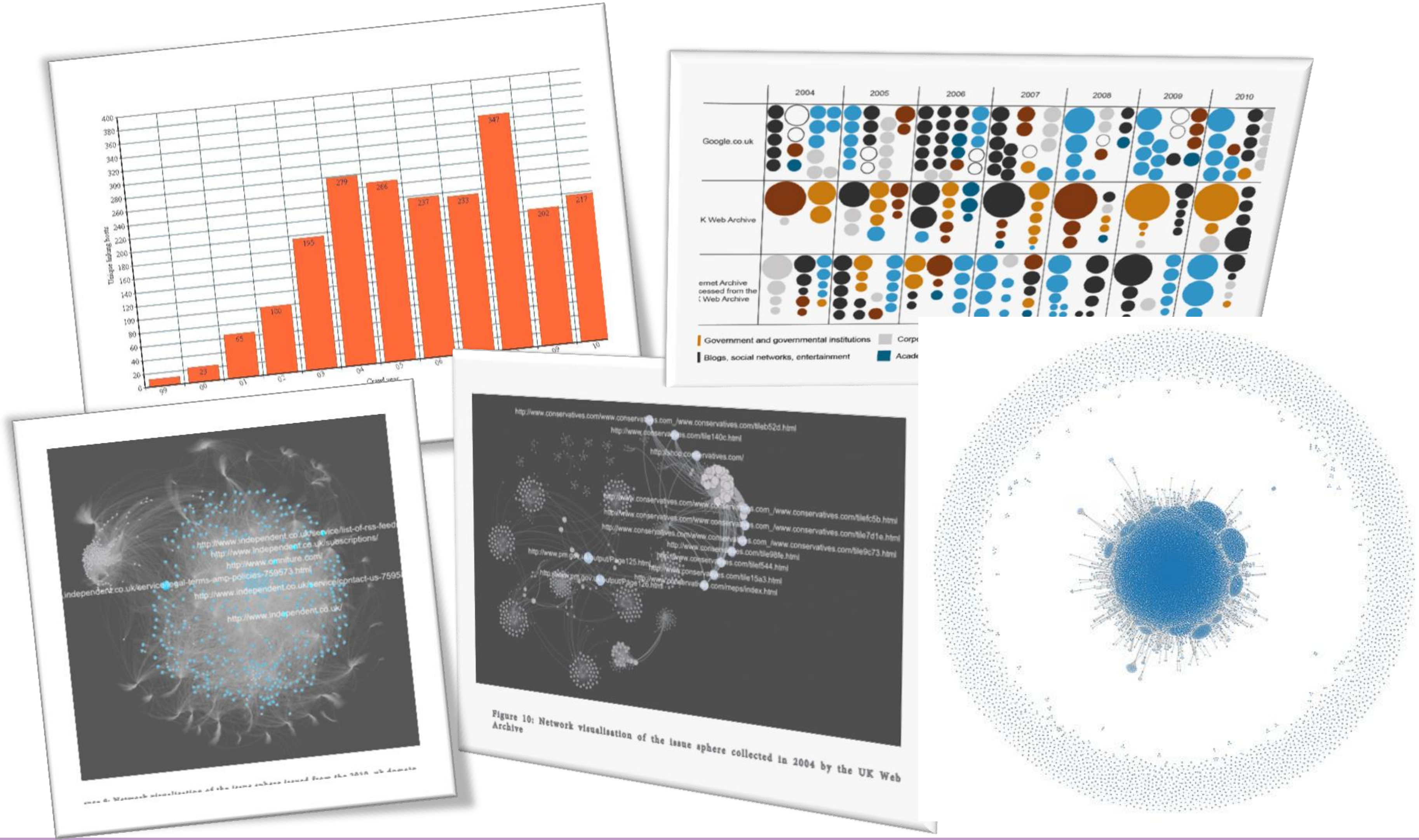
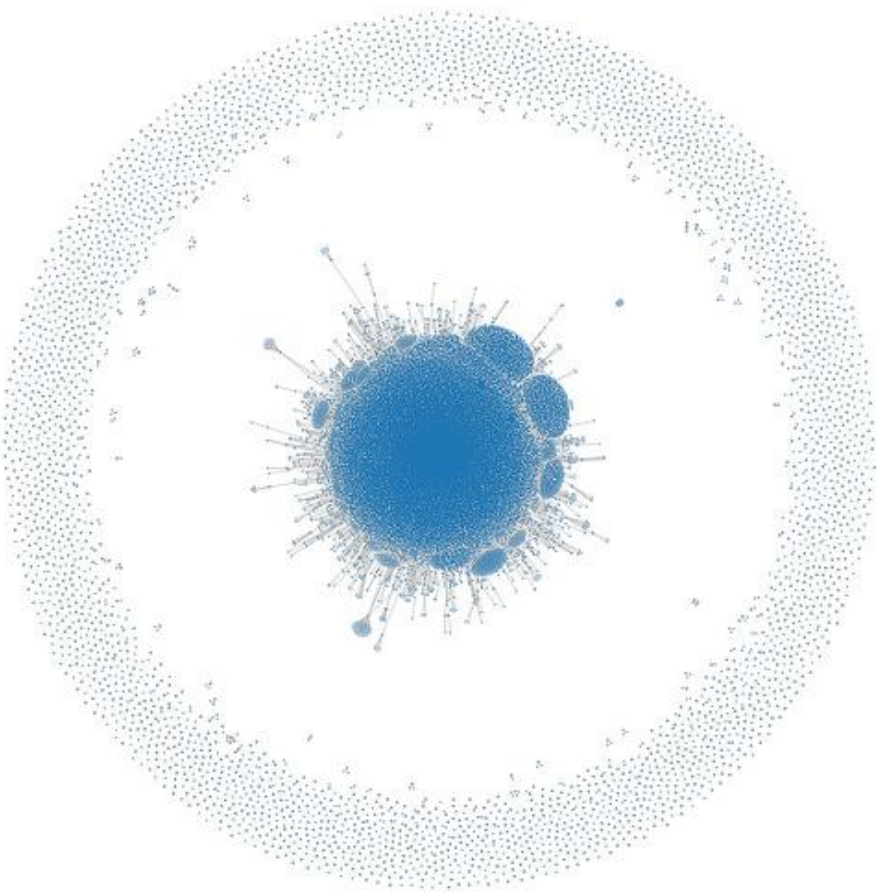


Figure 10: Network visualisation of the issue sphere collected in 2004 by the UK Web Archive

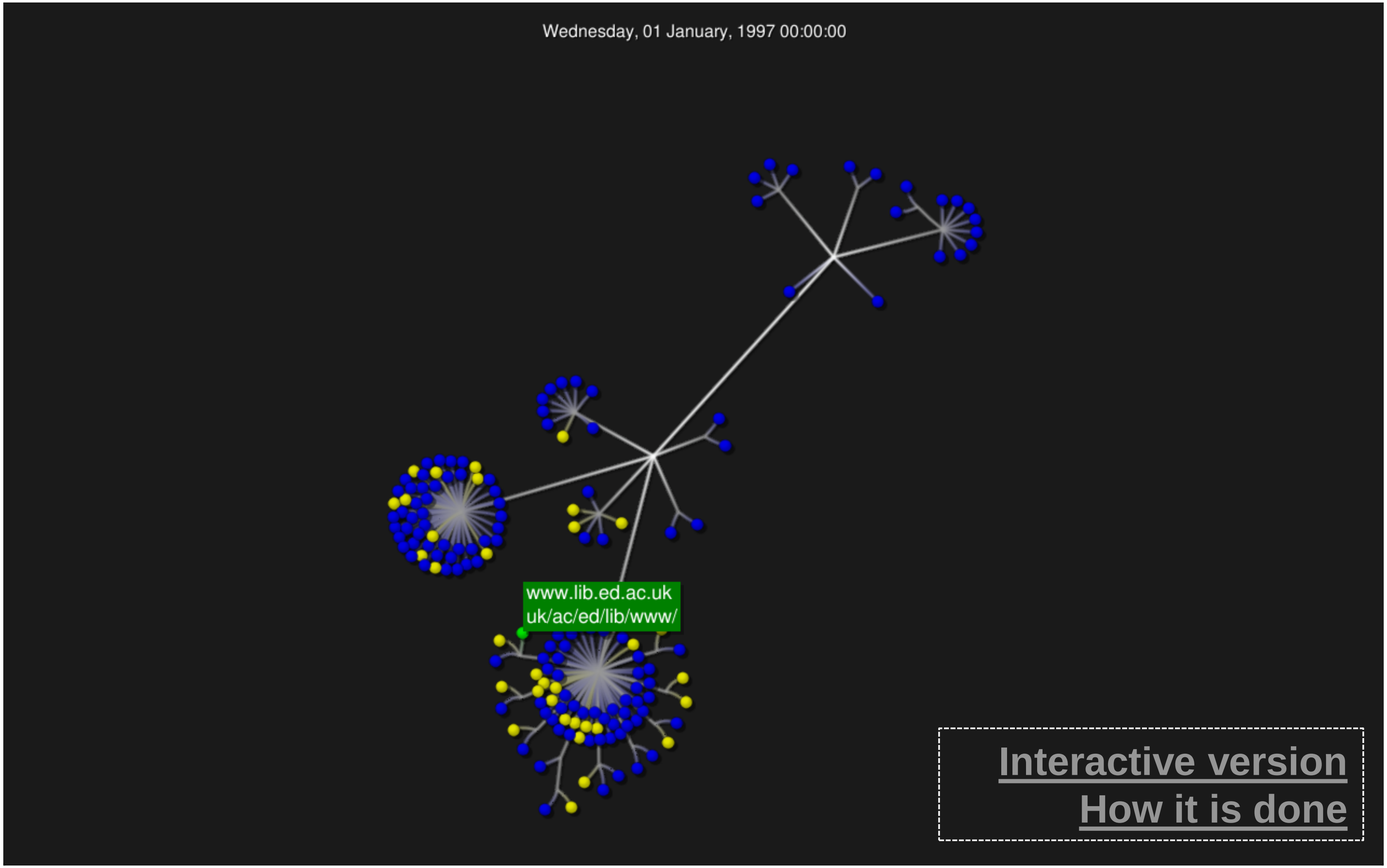
Courtesy of Peter Webster, Rainer Simon and Jules Mataly

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

Visualising links (to and from bl.uk)

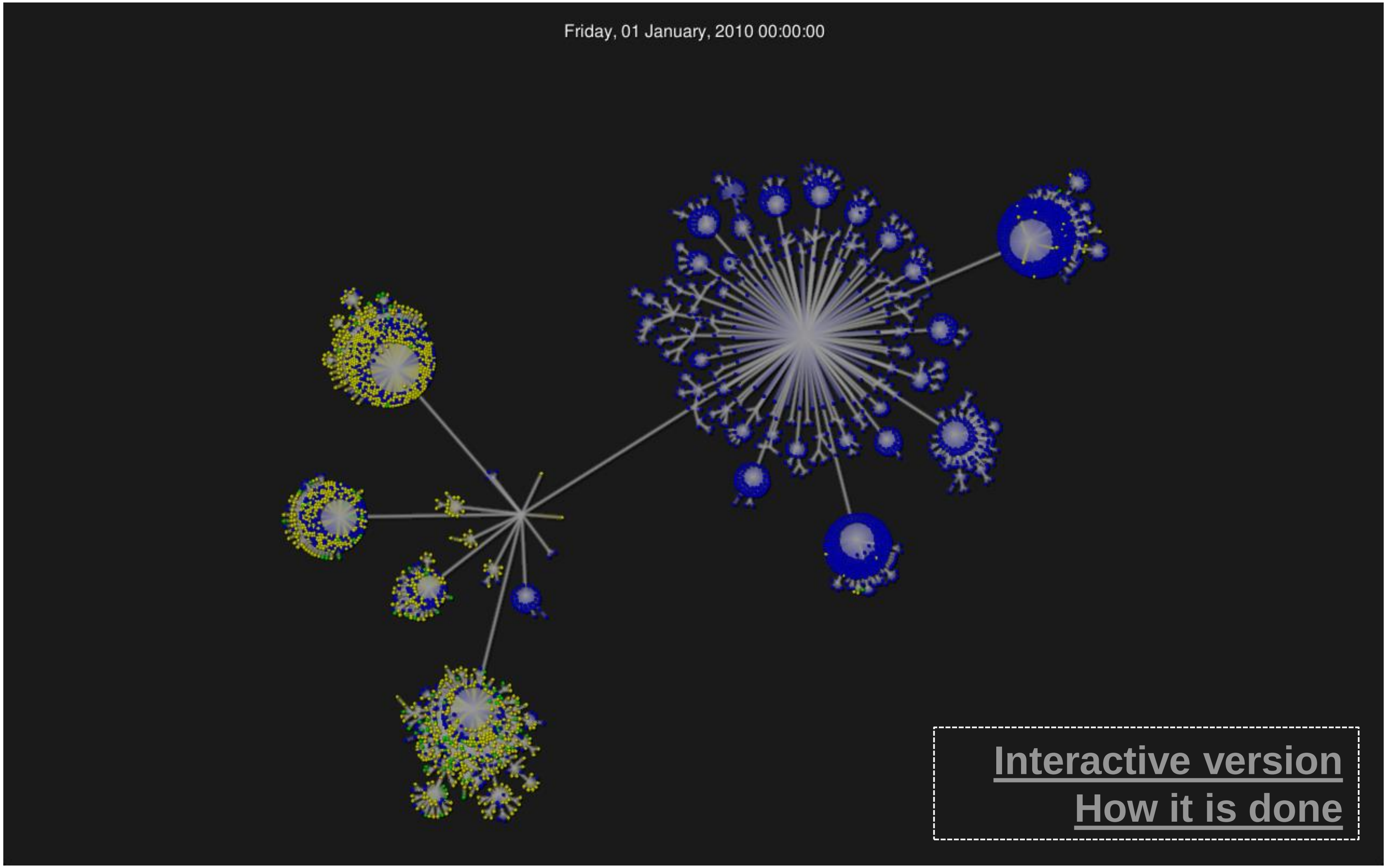


Wednesday, 01 January, 1997 00:00:00

Interactive version

How it is done

Visualising links (to and from bl.uk)

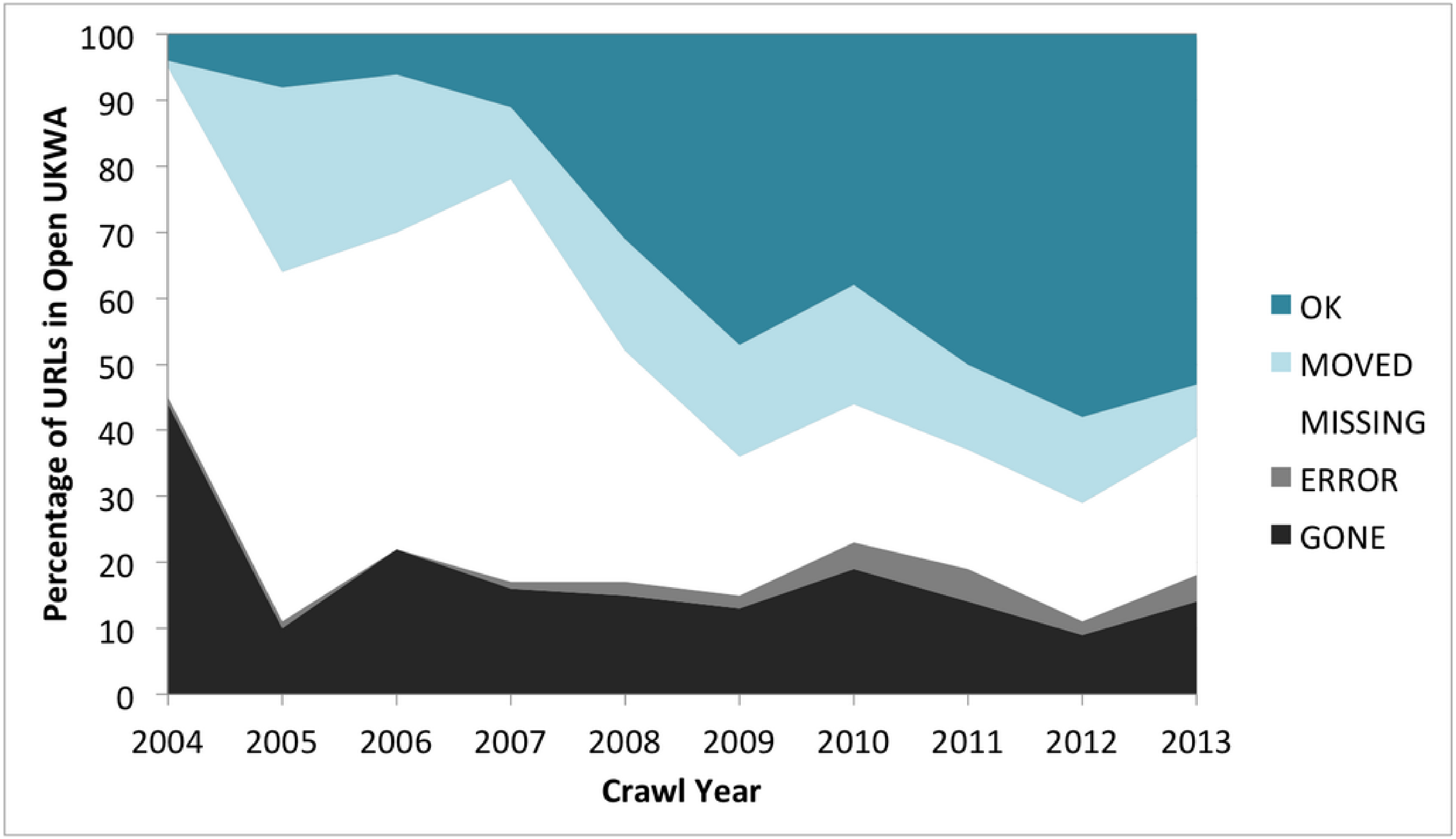


Friday, 01 January, 2010 00:00:00

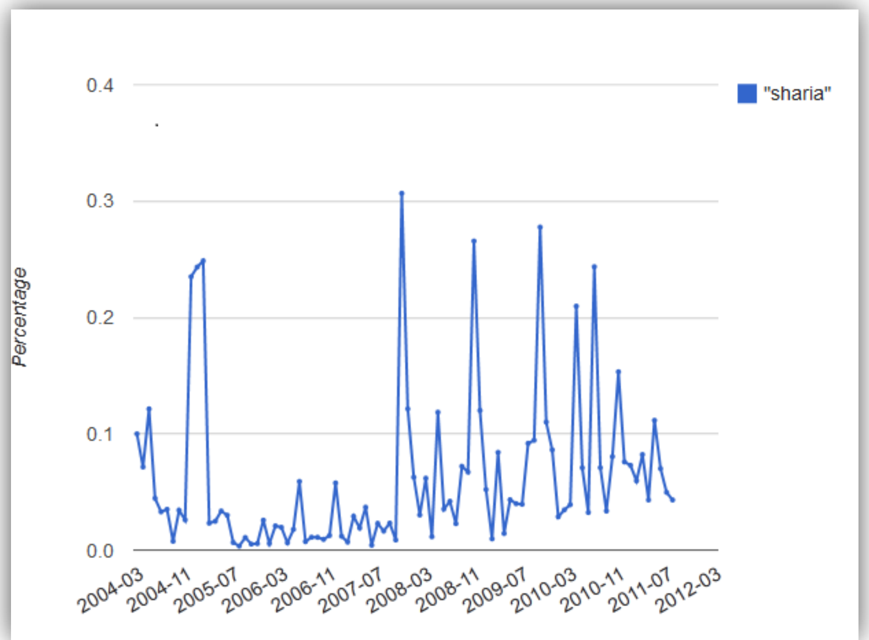
Interactive version

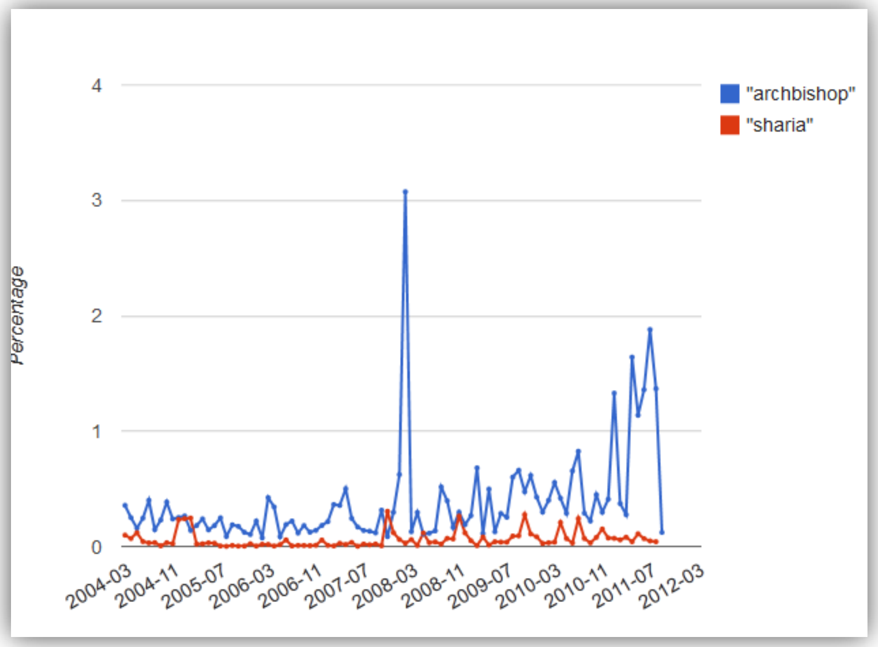
How it is done

The disappearing web over time



Using N-gram for scholarly research



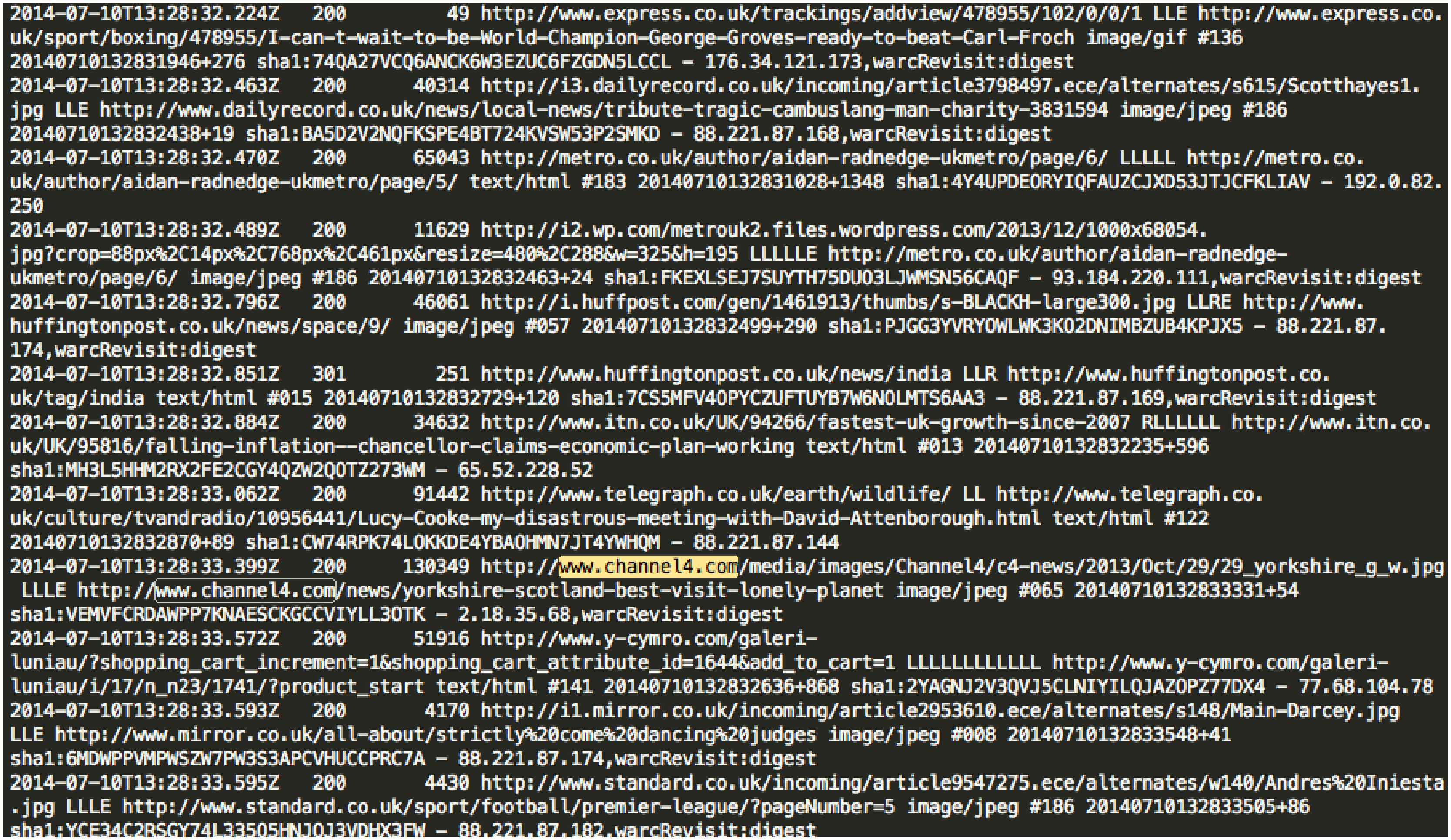


- *Courtesy of Dr Peter Webster, Institute of Historical Research, University of London*

Viral content, null results in the index

- 157,846 viral records detected in 2013 UK domain crawl
- Not included in the index
  - Redirects
  - Content no longer there (404)
  - Content there, but restricted
  - Server error

Crawl logs



```
2014-07-10T13:28:32.224Z   200         49 http://www.express.co.uk/trackings/addview/478955/102/0/0/1 LLE http://www.express.co.
uk/sport/boxing/478955/I-can-t-wait-to-be-World-Champion-George-Groves-ready-to-beat-Carl-Froch image/gif #136
20140710132831946+276 sha1:74QA27VCQ6ANCK6W3EZUC6FZGDN5LCCL - 176.34.121.173,warcRevisit:digest
2014-07-10T13:28:32.463Z   200      40314 http://i3.dailyrecord.co.uk/incoming/article3798497.ece/alternates/s615/Scotthayes1.
jpg LLE http://www.dailyrecord.co.uk/news/local-news/tribute-tragic-cambuslang-man-charity-3831594 image/jpeg #186
20140710132832438+19 sha1:BA5D2V2NQFKSPE4BT724KVSW53P2SMKD - 88.221.87.168,warcRevisit:digest
2014-07-10T13:28:32.470Z   200      65043 http://metro.co.uk/author/aidan-radnedge-ukmetro/page/6/ LLLLL http://metro.co.
uk/author/aidan-radnedge-ukmetro/page/5/ text/html #183 20140710132831028+1348 sha1:4Y4UPDEORYIQFAUZCJXD53JTJCFKLIAV - 192.0.82.
250
2014-07-10T13:28:32.489Z   200      11629 http://i2.wp.com/metrouk2.files.wordpress.com/2013/12/1000x68054.
jpg?crop=88px%2C14px%2C768px%2C461px&resize=480%2C288&w=325&h=195 LLLLLE http://metro.co.uk/author/aidan-radnedge-
ukmetro/page/6/ image/jpeg #186 20140710132832463+24 sha1:FKEXLSEJ7SUYTH75DUO3LJWMSN56CAQF - 93.184.220.111,warcRevisit:digest
2014-07-10T13:28:32.796Z   200      46061 http://i.huffpost.com/gen/1461913/thumbs/s-BLACKH-large300.jpg LLRE http://www.
huffingtonpost.co.uk/news/space/9/ image/jpeg #057 20140710132832499+290 sha1:PJGG3YVRYOWLWK3KO2DNIMBZUB4KPJX5 - 88.221.87.
174,warcRevisit:digest
2014-07-10T13:28:32.851Z   301        251 http://www.huffingtonpost.co.uk/news/india LLR http://www.huffingtonpost.co.
uk/tag/india text/html #015 20140710132832729+120 sha1:7CS5MFV4OPYCZUFTUYB7W6NOLMTS6AA3 - 88.221.87.169,warcRevisit:digest
2014-07-10T13:28:32.884Z   200      34632 http://www.itn.co.uk/UK/94266/fastest-uk-growth-since-2007 RLLLLLL http://www.itn.co.
uk/UK/95816/falling-inflation--chancellor-claims-economic-plan-working text/html #013 20140710132832235+596
sha1:MH3L5HHM2RX2FE2CGY4QZW2QOTZ273WM - 65.52.228.52
2014-07-10T13:28:33.062Z   200      91442 http://www.telegraph.co.uk/earth/wildlife/ LL http://www.telegraph.co.
uk/culture/tvandradio/10956441/Lucy-Cooke-my-disastrous-meeting-with-David-Attenborough.html text/html #122
20140710132832870+89 sha1:CW74RPK74LOKKDE4YBAOHMN7JT4YWHQM - 88.221.87.144
2014-07-10T13:28:33.399Z   200     130349 http://www.channel4.com/media/images/Channel4/c4-news/2013/Oct/29/29_yorkshire_g_w.jpg
 LLLE http://www.channel4.com/news/yorkshire-scotland-best-visit-lonely-planet image/jpeg #065 20140710132833331+54
sha1:VEMVFCRDAWPP7KNAESCKGCCVIYLL3OTK - 2.18.35.68,warcRevisit:digest
2014-07-10T13:28:33.572Z   200      51916 http://www.y-cymro.com/galeri-
luniau/?shopping_cart_increment=1&shopping_cart_attribute_id=1644&add_to_cart=1 LLLLLLLLLLLL http://www.y-cymro.com/galeri-
luniau/i/17/n_n23/1741/?product_start text/html #141 20140710132832636+868 sha1:2YAGNJ2V3QVJ5CLNIYILQJAZOPZ77DX4 - 77.68.104.78
2014-07-10T13:28:33.593Z   200       4170 http://i1.mirror.co.uk/incoming/article2953610.ece/alternates/s148/Main-Darcey.jpg
LLE http://www.mirror.co.uk/all-about/strictly%20come%20dancing%20judges image/jpeg #008 20140710132833548+41
sha1:6MDWPPVMPWSZW7PW3S3APCVHUCCPRC7A - 88.221.87.174,warcRevisit:digest
2014-07-10T13:28:33.595Z   200       4430 http://www.standard.co.uk/incoming/article9547275.ece/alternates/w140/Andres%20Iniesta
.jpg LLLE http://www.standard.co.uk/sport/football/premier-league/?pageNumber=5 image/jpeg #186 20140710132833505+86
sha1:YCE34C2RSGY74L33505HNJOJ3VDHX3FW - 88.221.87.182,warcRevisit:digest
```

Next steps

- General issues related to analytical access
  - Scepticism or suspicion about hidden algorithms behind analysis
  - Biases in data and how data collection decisions lead to variances in outputs
  - Need to manage expectations, analysis and visualisation as finished products and first steps
  - Ethical and privacy issues

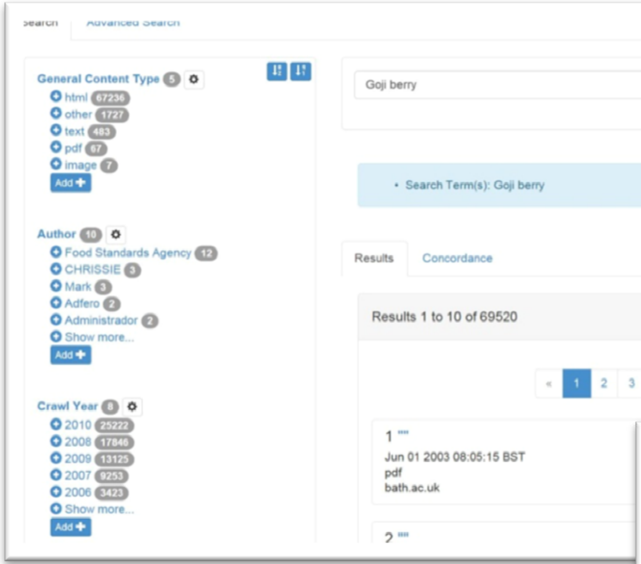
- Maximise transparency and make base-line knowledge is self-explanatory
  - e.g. scope of the archive, its coverage and lacunae, how it was collected, and how a particular website was crawled

- Further work through involvement in projects
  - Big UK Domain Data for Arts and Humanities
  - "ALEXANDRIA: Foundations for Temporal Retrieval, Exploration and Analytics in Web Archives
  - RESAW, a Research Infrastructure for the Study of Archived Web Materials

Big UK Domain Data for Arts and Humanities

- Funded by the UK Arts and Humanities Research Council as one of the 21 “Big Data” projects
- Collaboration between the 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
- A major study of the history of UK web space from 1996 to 2013 + sub-projects covering a range of disciplines
- Also an online training course and peer-reviewed journal articles.

Shine



Search | Advanced Search

General Content Type 5
- html 67236
- other 1727
- text 483
- pdf 67
- image 7

Add

Author 10
- Food Standards Agency 12
- CHRISSIE 3
- Mark 3
- Adfero 2
- Administrator 2
- Show more...

Add

Crawl Year 8
- 2010 25222
- 2008 17846
- 2009 13125
- 2007 9253
- 2006 3423
- Show more...

Add

Goji berry | Normal Mode | Search | Reset

- Search Term(s): Goji berry

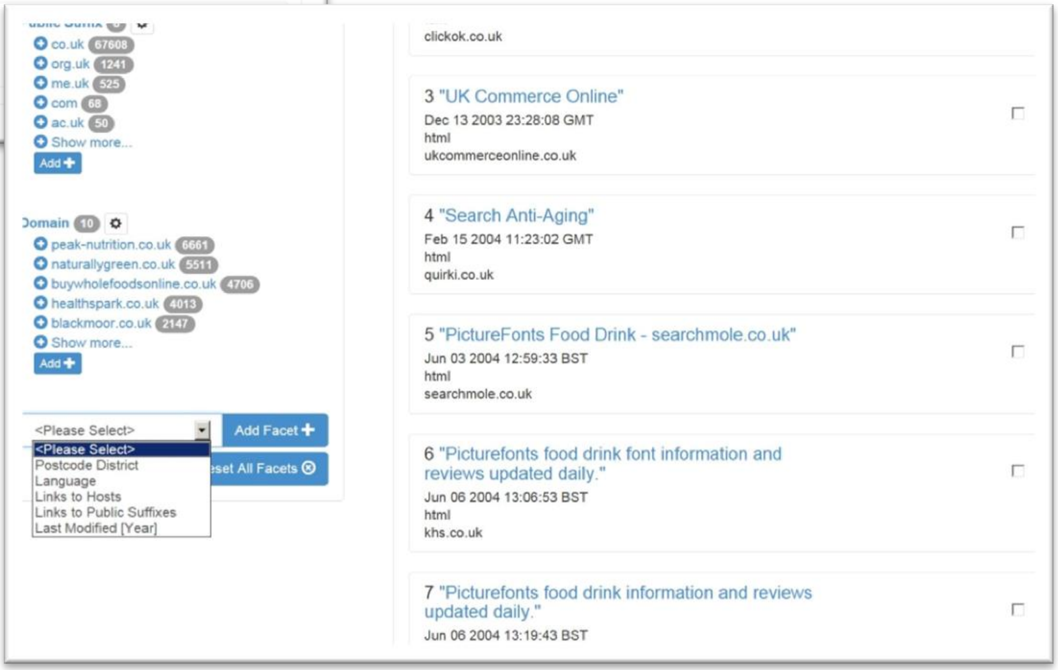
Results | Concordance

Results 1 to 10 of 69520 | CSV | Asc

« 1 2 3 4 5 6 »

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

2 ""



- co.uk 67608
- org.uk 1241
- me.uk 525
- com 68
- ac.uk 50
- Show more...

Add

Domain 10
- peak-nutrition.co.uk 6661
- naturallygreen.co.uk 5511
- buywholefoodsonline.co.uk 4706
- healthspark.co.uk 4013
- blackmoor.co.uk 2147
- Show more...

Add

<Please Select> | Add Facet
- <Please Select>
- Postcode District
- Language
- Links to Hosts
- Links to Public Suffixes
- Last Modified [Year]

Reset All Facets

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

- Brügger, N.; Finnemann, N.E., The Web and Digital Humanities: Theoretical and Methodological Concerns http://thelecturn.com/wp-content/uploads/2013/07/The-web-and-digital-humanities-Theoretical-and-Methodological-Concerns.pdf
- ISO technical report: Statistics and Quality Indicators for Web Archiving http://netpreserve.org/sites/default/files/resources/SO_TR_14873__E__2012-10-02_DRAFT.pdf
- UK Web Archive open data http://data.webarchive.org.uk/opendata/