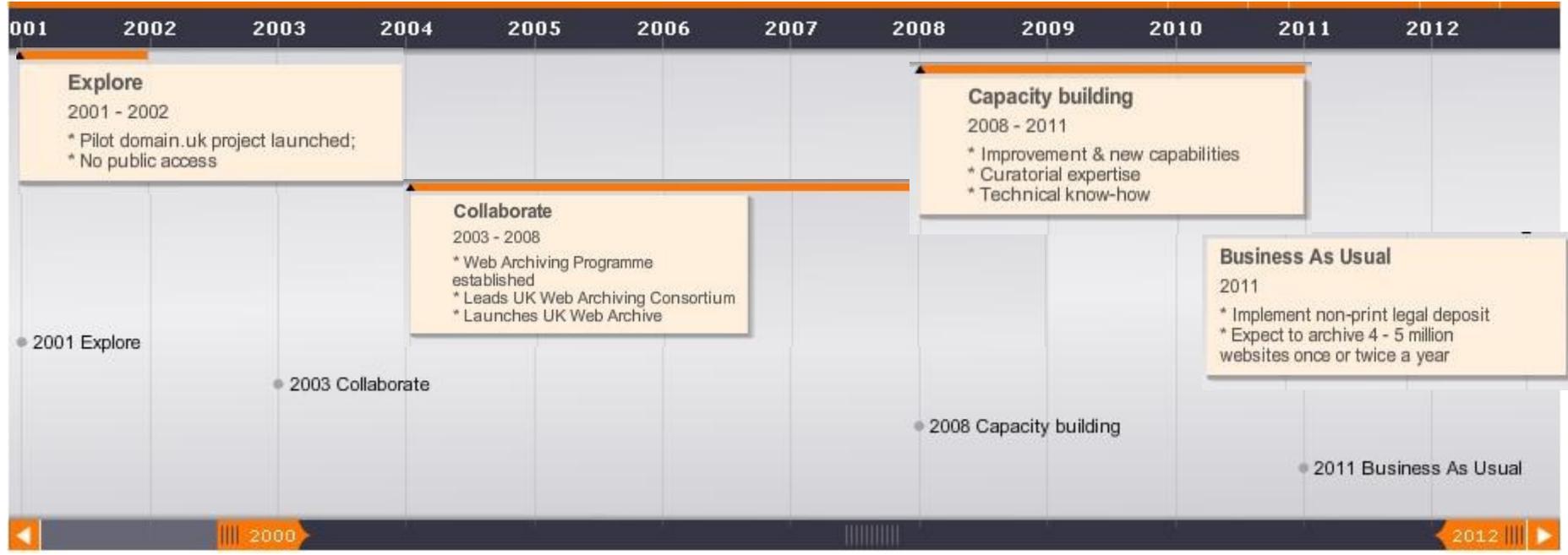


Analytical Access to the UK Web Archive: *data mining with intent*

Maureen Pennock,
Web Archive Engagement & Liaison Manager

Lewis Crawford,
Web Archive Technical Lead

UK Web Archive: Growing from strength to strength



Legacy access

UK Web Archive interface, June 2005



UK WEB ARCHIVING
CONSORTIUM
www.webarchive.org.uk

Topic Help **Webmaster Information**

Subjects Menu:

- About UK Web Archive
- Press Releases
- Consortium Partners
- Privacy Statement
- Copyright
- Contact Us
- Topic Help
- Webmaster Information
- Submission Form

Arts & Humanities	Government & Politics	Reference Works
Business & Economy	Health	Science & Technology
Education & Research	News & Media	Society & Culture

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

Modern access

Contemporary interface (2011): homepage

Translate to Welsh



You are here: Home

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

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.

- ### Quick website links
- What is the UK Web Archive?
 - Who is the UK Web Archive for?
 - How do I search the archive?
 - How can I nominate a website?

Quick search

Please enter text

Title (for a specific archived website)

Full text (across all the archived websites)

[Advanced search](#)

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).

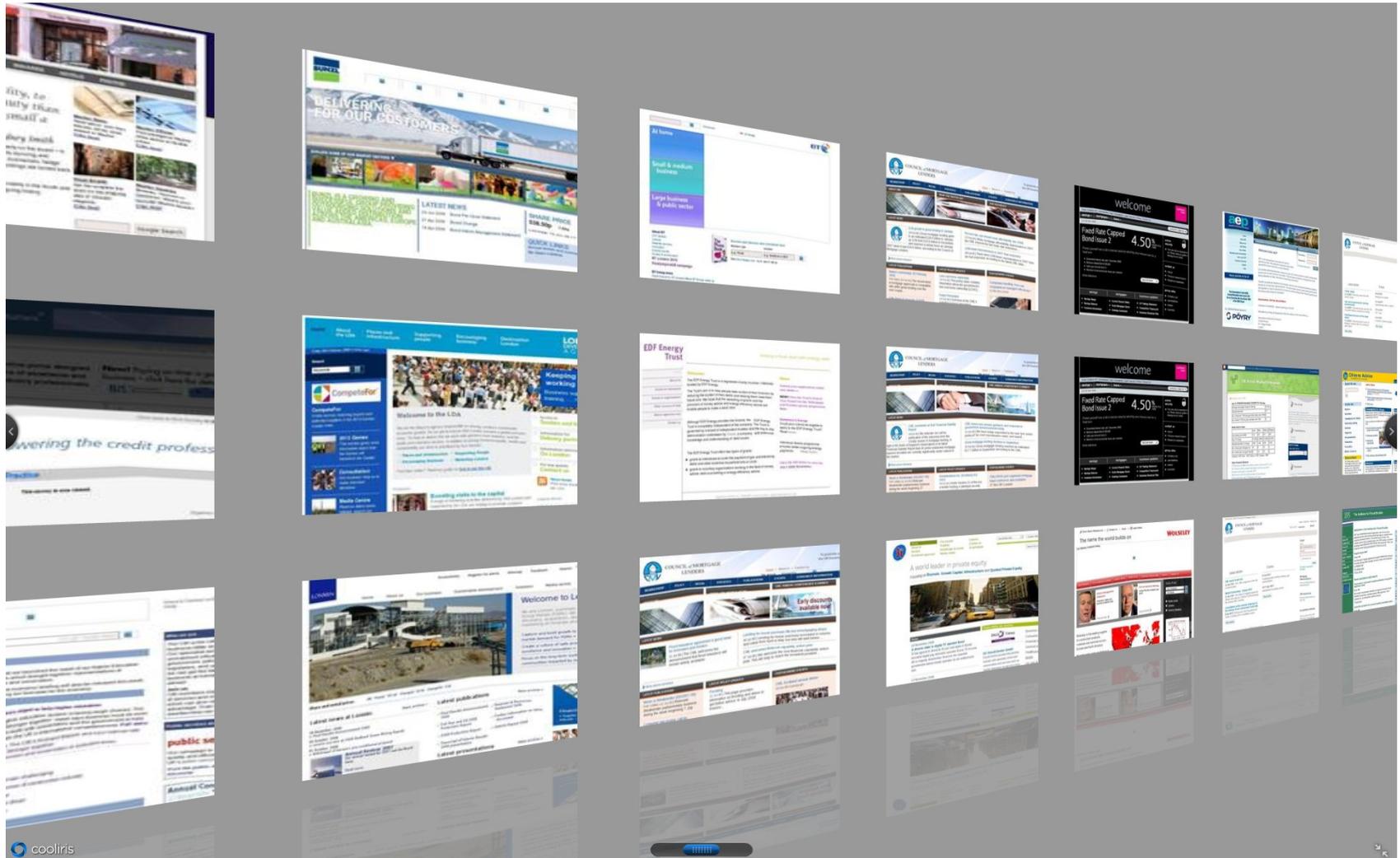


- ### Browse by Subject
- Education & Research
 - Arts & Humanities
 - Science & Technology
 - Government, Law & Politics
 - Society & Culture
 - Business, Economy & Industry
 - Medicine & Health
- [Browse by Subject](#)

Initiated by:

Modern access

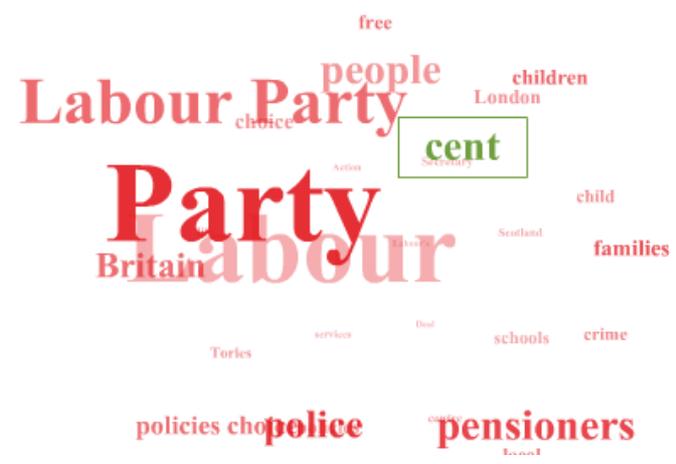
Contemporary interface
(2011): 3D Wall



Modern access

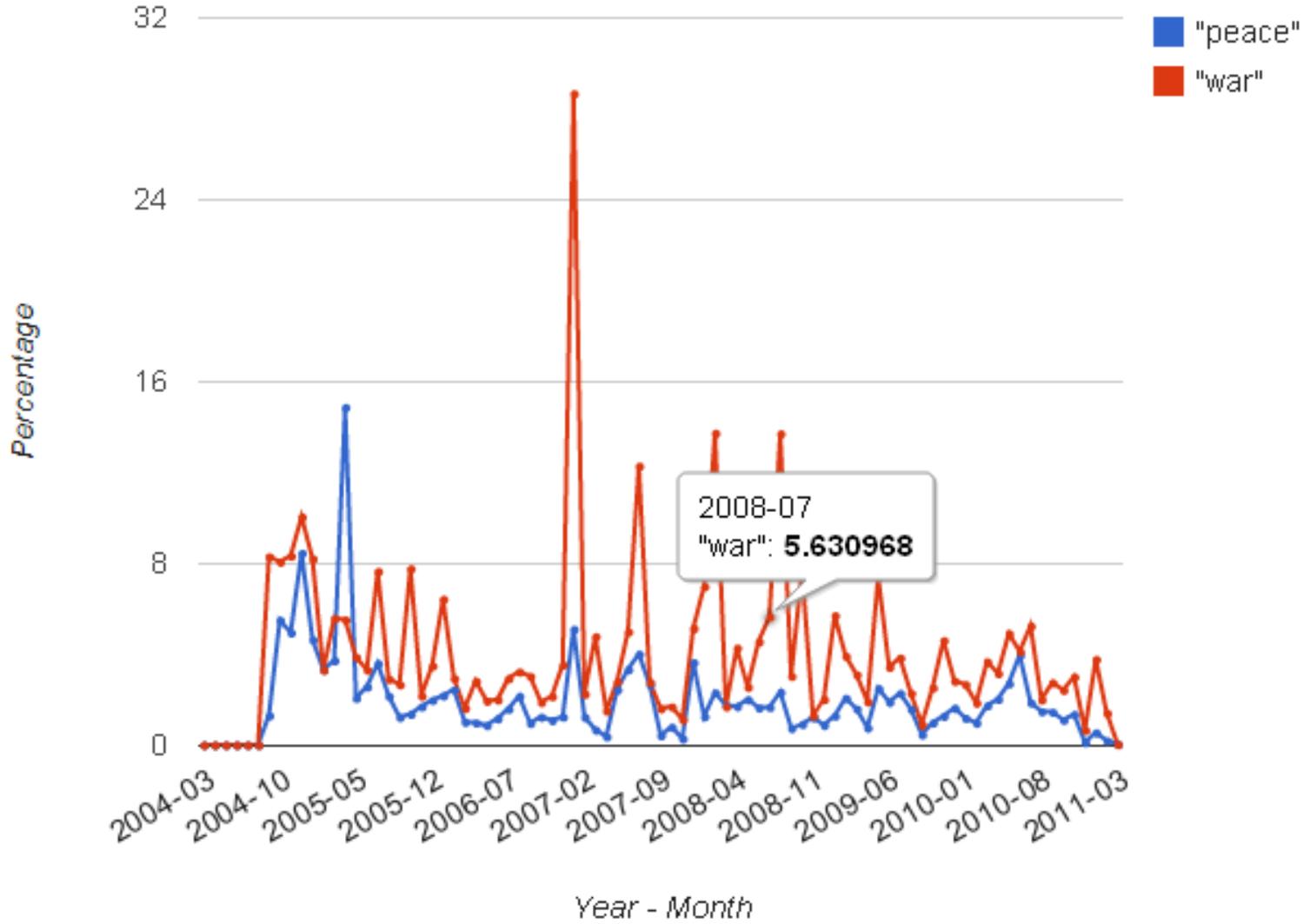
Contemporary interface:
WordClouds

- Special Collection 2005 general election
 - 147 websites archived during and immediately after the UK general election campaign of 2005.
 - Tag clouds (or weighted lists) generated for websites belonging to key political parties
 - Shows the most frequently used words in the websites during the 2005 election campaign
- Special collection 2010 general election now available



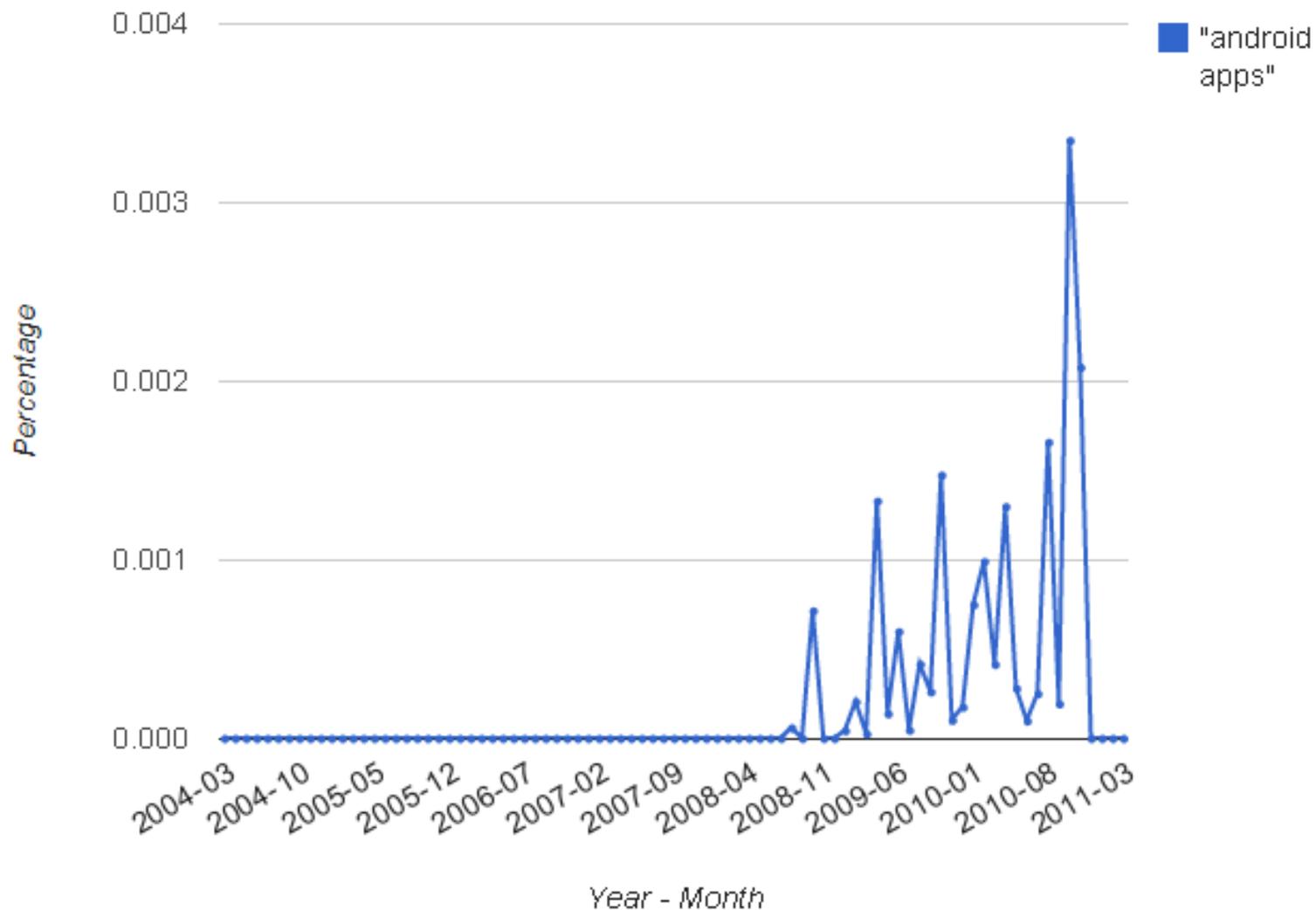
Modern access

Current interface: Ngram data visualisations



Modern access

Current interface: Ngram data visualisations



N-grams

The screenshot shows the UK Web Archive N-gram Search page. At the top, there is a navigation bar with the UK Web Archive logo and a row of thumbnails for various archived dates. Below this is a breadcrumb trail: "You are here: > UK Web Archive N-gram".

The main content area is titled "UK Web Archive N-gram Search". It includes a brief explanation: "When you enter terms / phrases into the search box, it displays a graph showing how these phrases have occurred in the UK Web Archive over time (e.g., 'London', 'Paris' or 'Tony Blair', 'Gordon Brown', 'David Cameron'). Clicking on a specific value will display the actual search results for that term/phrase."

Below the text is a search interface with a text input field containing the text "Enter words or phrases separated by commas", a "Generate N-gram!" button, and a line graph. The graph plots the percentage of occurrences for three terms: "Gordon Brown" (blue line), "Tony Blair" (red line), and "David Cameron" (orange line) from March 2004 to March 2011. The Y-axis is labeled "Percentage" and ranges from 0 to 20. The X-axis is labeled "Year - Month" and shows dates from 2004-03 to 2011-03. A significant peak is visible for all three terms around late 2008, with "Tony Blair" reaching approximately 16%.

At the bottom of the page, there are links for "Notice and takedown", "Terms and conditions", and "Privacy statement".

Media based Results

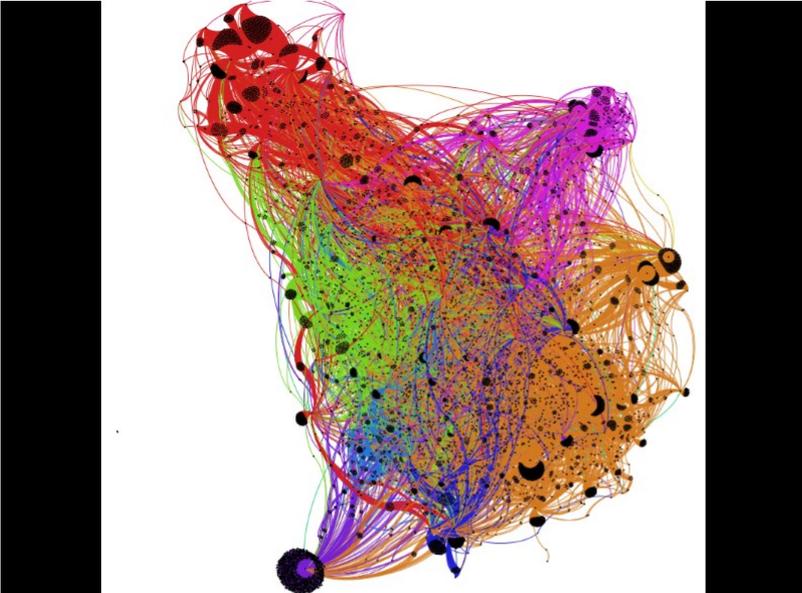
The screenshot shows a web browser window displaying the UK Web Archive search results for the term 'Cameron'. The browser's address bar shows the URL: `mosaic.bl.uk/ukwa/search/page/1?text=Cameron&option_search=image`. The page features a navigation menu on the left, a main search results area with tabs for 'Document Results', 'Image Results', and 'Media Results', and a right-hand sidebar with filters for 'Date Ranges' and 'Select a Collection'. The search results are displayed in two sections, both showing 1238 results. The first section shows a grid of image thumbnails, including several portraits of David Cameron and a 'King' logo. The second section shows a grid of video thumbnails, including a scene with a 'POLICE' sign. The page footer contains links for 'Notice and takedown', 'Terms and conditions', and 'Privacy statement'.

What is in the web archive?

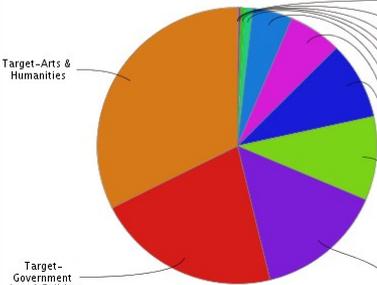
UK Web Archive Gephi Seadragon Export

www.webarchive.org.uk/ukwa2/seadragon.html

Gephi Seadragon Export



Key



- Sub-Group
- Target-Religion
- Event
- Collection
- Subject
- Target-Science & Technology
- Target-Medicine & Health
- Target-Society & Culture
- Target-Business Economy & Industry
- Target-Education & Research
- Target-Government Law & Politics
- Target-Arts & Humanities

UK Web Archive Gephi Seadragon Export

www.webarchive.org.uk/ukwa2/seadragon.html

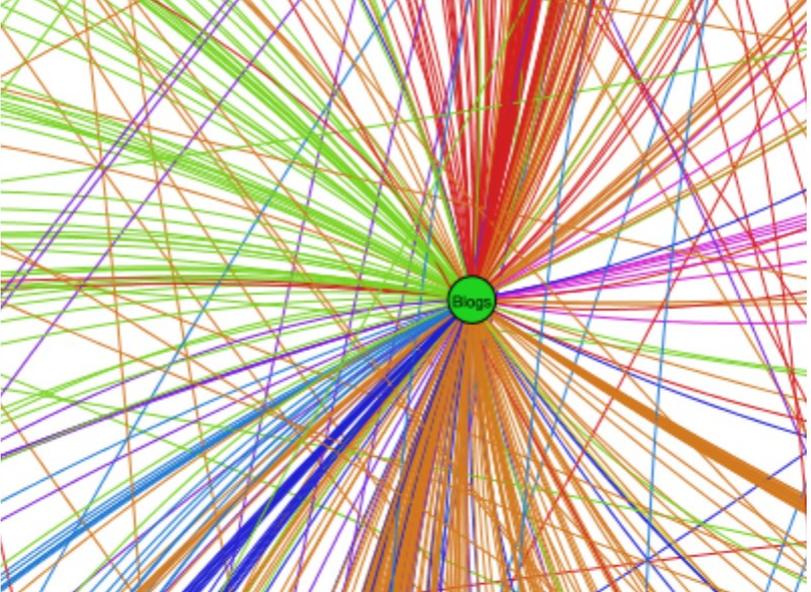
Gephi Seadragon Export



UK Web Archive Gephi Seadragon Export

www.webarchive.org.uk/ukwa2/seadragon.html

Gephi Seadragon Export



Semantic Analysis

The screenshot shows a browser window displaying a search result from the UK Web Archive. The search query is for 'Cameron' and the result is for the 'Conservative Party' page. A semantic analysis overlay is present, providing structured information about the page's content.

Language
 english [Ethnologue](#) [Wikipedia](#)

Category
 culture_politics

Keywords
 David Cameron, Conservative, local conservatives, Conservative Party, Cameron appeals,

Concepts

Conservative Party
 Website: <http://www.conservatives.com/>
 DBPEDIA: [http://dbpedia.org/resource/Conservative_Party_\(UK\)](http://dbpedia.org/resource/Conservative_Party_(UK))
 OPENCYC: http://sw.opencyc.org/concept/Mx4rPn-1Pq2iQdId_uIm5v7paw
 FREEBASE: <http://rdf.freebase.com/ns/quad.9202a8c04000641f80000000003eeb3>

United Kingdom general election, 2005
 DBPEDIA: http://dbpedia.org/resource/United_Kingdom_general_election_2005
 FREEBASE: <http://rdf.freebase.com/ns/quad.9202a8c04000641f8000000000605a7f>

Conservatism
 DBPEDIA: <http://dbpedia.org/resource/Conservatism>
 OPENCYC: <http://sw.opencyc.org/concept/Mx4rvXQ3x1wpEbGdrcN5Y29vcA>
 FREEBASE: <http://rdf.freebase.com/ns/quad.9202a8c04000641f80000000000f6d5>

John Major
 Website: <http://www.johnmajor.co.uk/>
 DBPEDIA: http://dbpedia.org/resource/John_Major
 OPENCYC: <http://sw.opencyc.org/concept/Mx4rvnGrT5wpEbGdrcN5Y29vcA>
 FREEBASE: <http://rdf.freebase.com/ns/quad.9202a8c04000641f80000000000206a7>

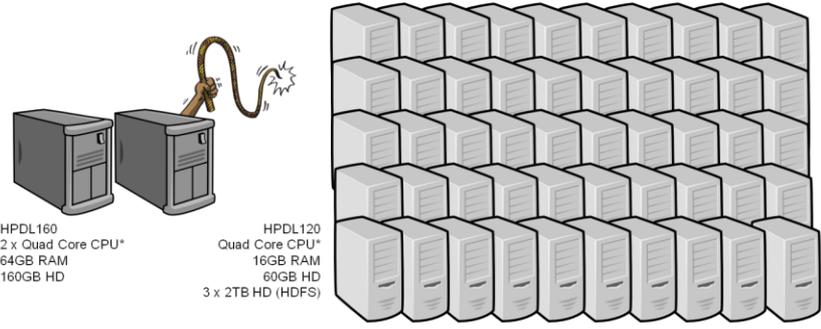
Leader of the Opposition
 DBPEDIA: [http://dbpedia.org/resource/Leader_of_the_Opposition_\(United_Kingdom\)](http://dbpedia.org/resource/Leader_of_the_Opposition_(United_Kingdom))
 FREEBASE: <http://rdf.freebase.com/ns/quad.9202a8c04000641f80000000000326bc8>

Entities Summary	Entities Detail
Persons 4	David Cameron Name: David Cameron Website: http://www.davidcameronmp.com/ DBPEDIA: http://dbpedia.org/resource/David_Cameron FREEBASE: http://rdf.freebase.com/ns/quad.9202a8c04000641f8000000000231a4a
Blair 3	Blair Name: Tony Blair Website: http://www.tonyblairoffice.org/ DBPEDIA: http://dbpedia.org/resource/Tony_Blair OPENCYC: http://sw.opencyc.org/concept/Mx4rvalD55wpEbGdrcN5Y29vcA FREEBASE: http://rdf.freebase.com/ns/quad.9202a8c04000641f8000000000922391
Jobs 1	
Alan Mabbutt 1	
City 1	
London 1	
Country 1	
Policy & Campaigns 6	

<http://www.conservatives.com/tile7ddb.html>
 Instance Date: 2006-06-15
 All Dates:
 Conservative Party - David Cameron - News Text size: A A A | Accessibility | Site map | Search the site: | Advanced Search Find your local Conservatives Join our mailing list

Big Data Analytics

- 52 servers:
 - 1 JobTracker
 - 1 NameNode
 - 50 DataNodes/TaskTrackers



UK Web Archive | Hue | bellie-private:8088

Hi Icraver for | Logout | Shortcuts

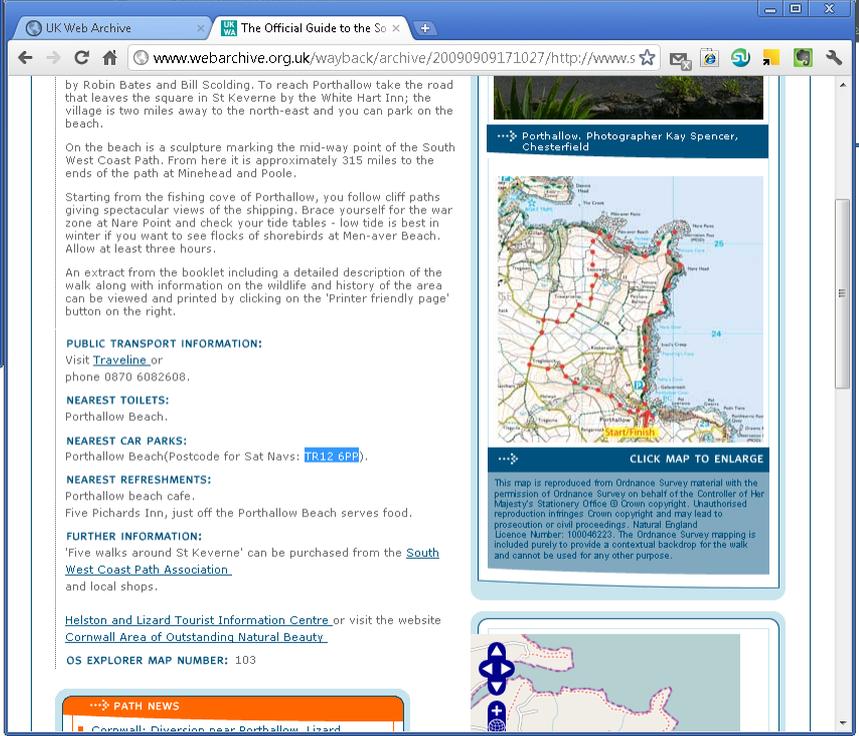
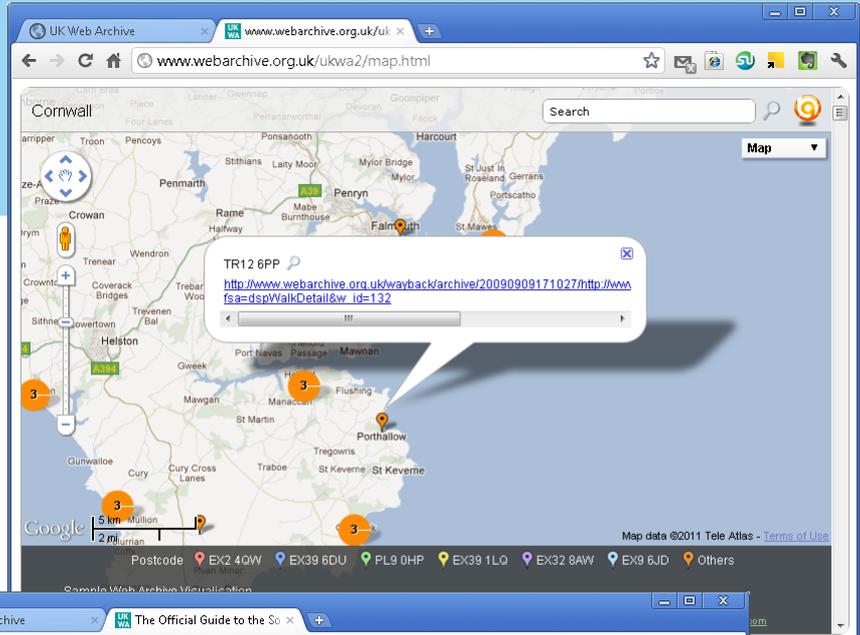
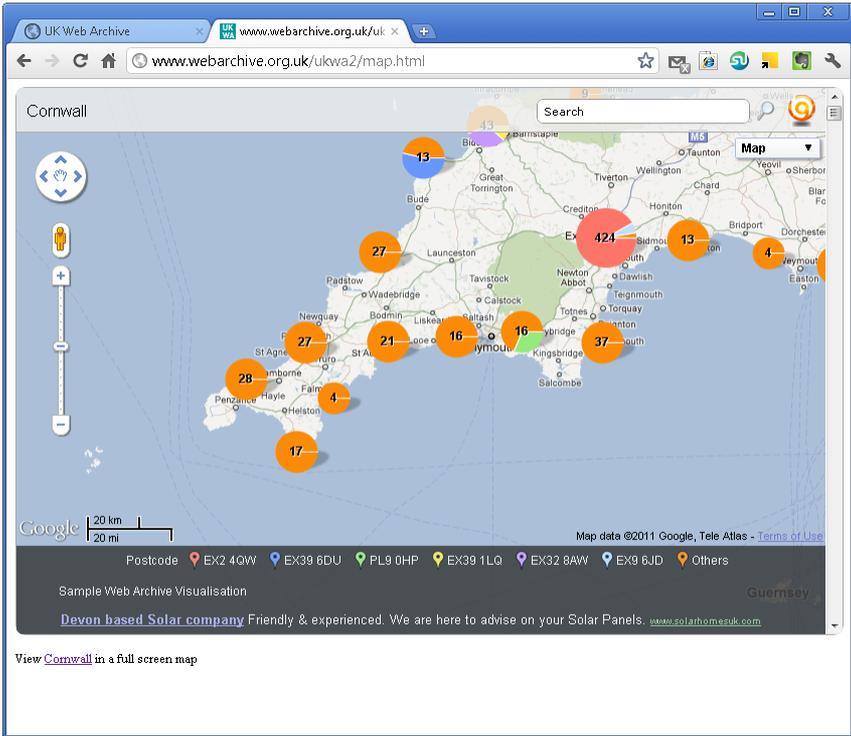
Beeswax Table Metadata: postcode

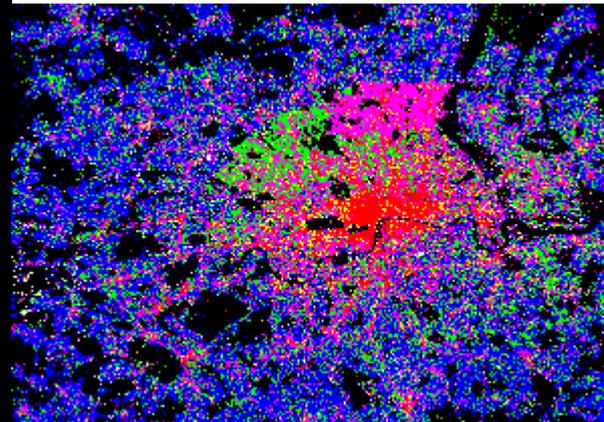
Query Editor | My Queries | Saved Queries | History | Tables | Settings

Sample	Columns
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/assets/wtd040141.rtf	postcode B15 3TB
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/assets/wtd040144.rtf	EC1A 7BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/contact.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/contact.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/contact.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038497.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038862.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038862.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038871.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038873.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038886.html	NW1 2BE
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	SW7 5BD
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	SW1V 2QQ
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	W1M 8AL
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	NW1 2DB
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	NW1 3DG
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	NW1 2DB
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	NW1 2DB
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	WC1H 9JP
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	WC2N 5NG
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	SE1 6LW
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	EC1R 1UW
2004-09-23T12:00:00Z/http://library.wellcome.ac.uk/doc_WTL038910.html	E1 7NT

feedback | selenium-server-stan...jar | Show all downloads...

Where is the web archive?





1:	Blue
2-5:	Green
5+	Purple
50+	Yellow
100+	Red

Thank you!

Digital (Data-> Information -> Knowledge) is a renewable resource and one that we can and should exploit.