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

001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012

Explore
2001 - 2002
* Pilot domain.uk project launched;
* No public access

Collaborate
2003 - 2008
* Web Archiving Programme established
* Leads UK Web Archiving Consortium
* Launches UK Web Archive

Capacity building
2008 - 2011
* Improvement & new capabilities
* Curatorial expertise
* Technical know-how

Business As Usual
2011
* Implement non-print legal deposit
* Expect to archive 4 - 5 million websites once or twice a year

2001 Explore
2003 Collaborate
2008 Capacity building
2011 Business As Usual

2000 2012

Legacy access

UK Web Archive interface, June 2005

Topic Help | Webmaster Information

Search | Subjects Menu: -- Select --

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

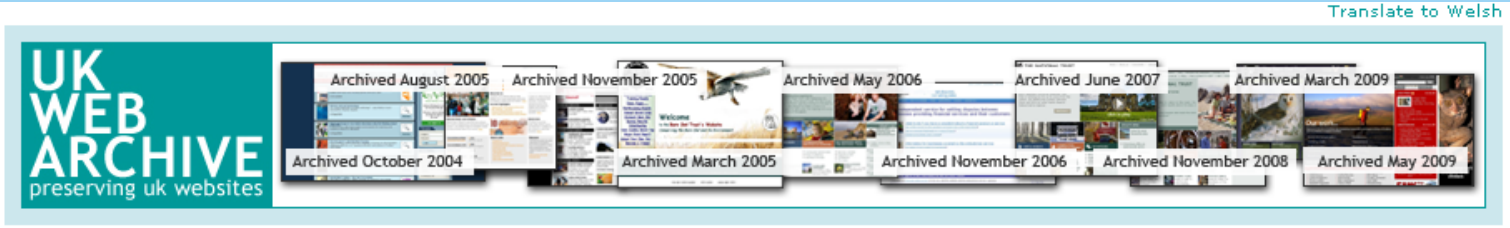
Arts & Humanities	Government & Politics	Reference Works
Business & Economy	Health	Science & Technology
Education & Research	News & Media	Society & Culture

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



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

Labour

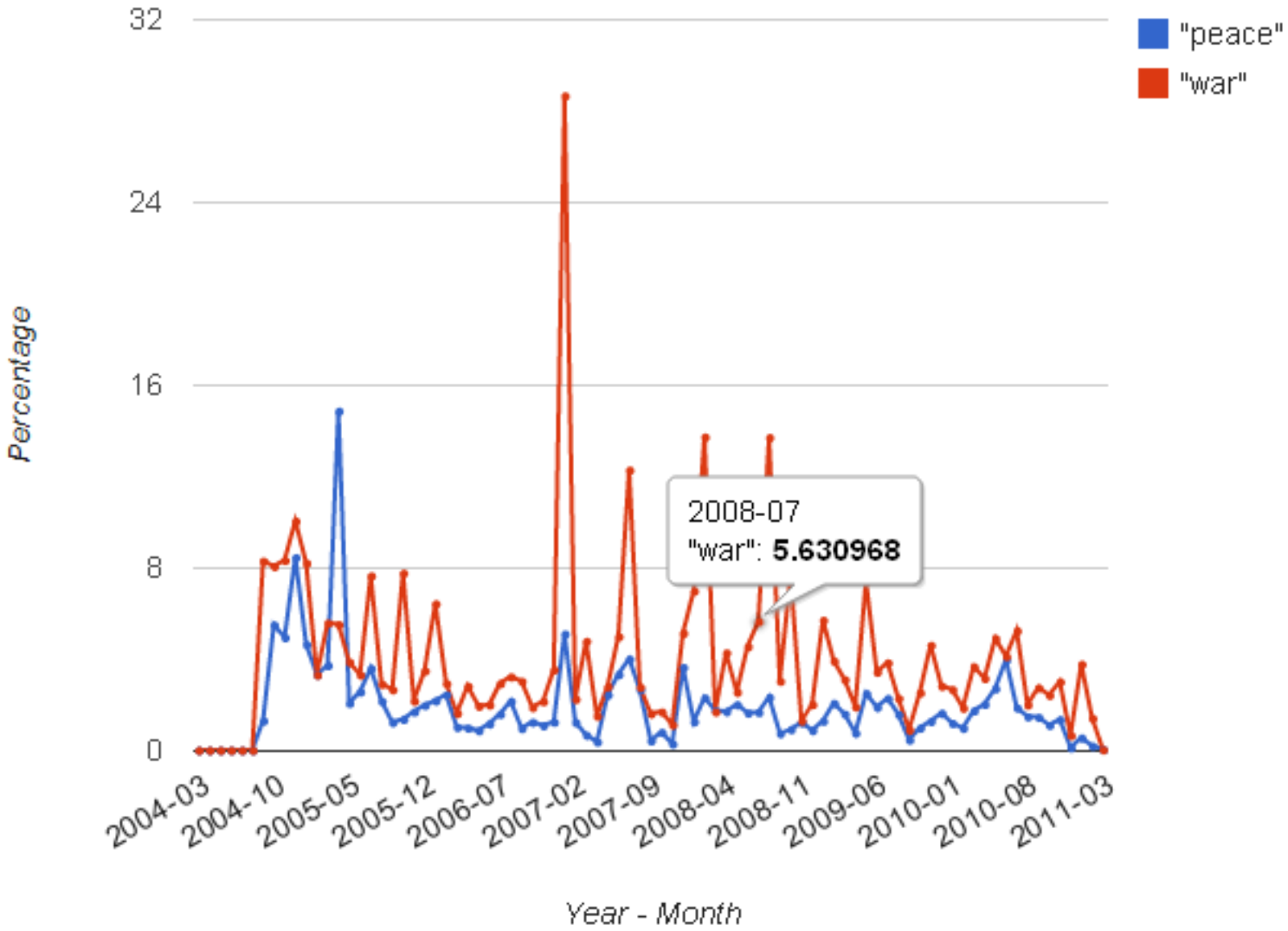
Conservative

Liberal Democrats

• Special collection 2010 general election now available

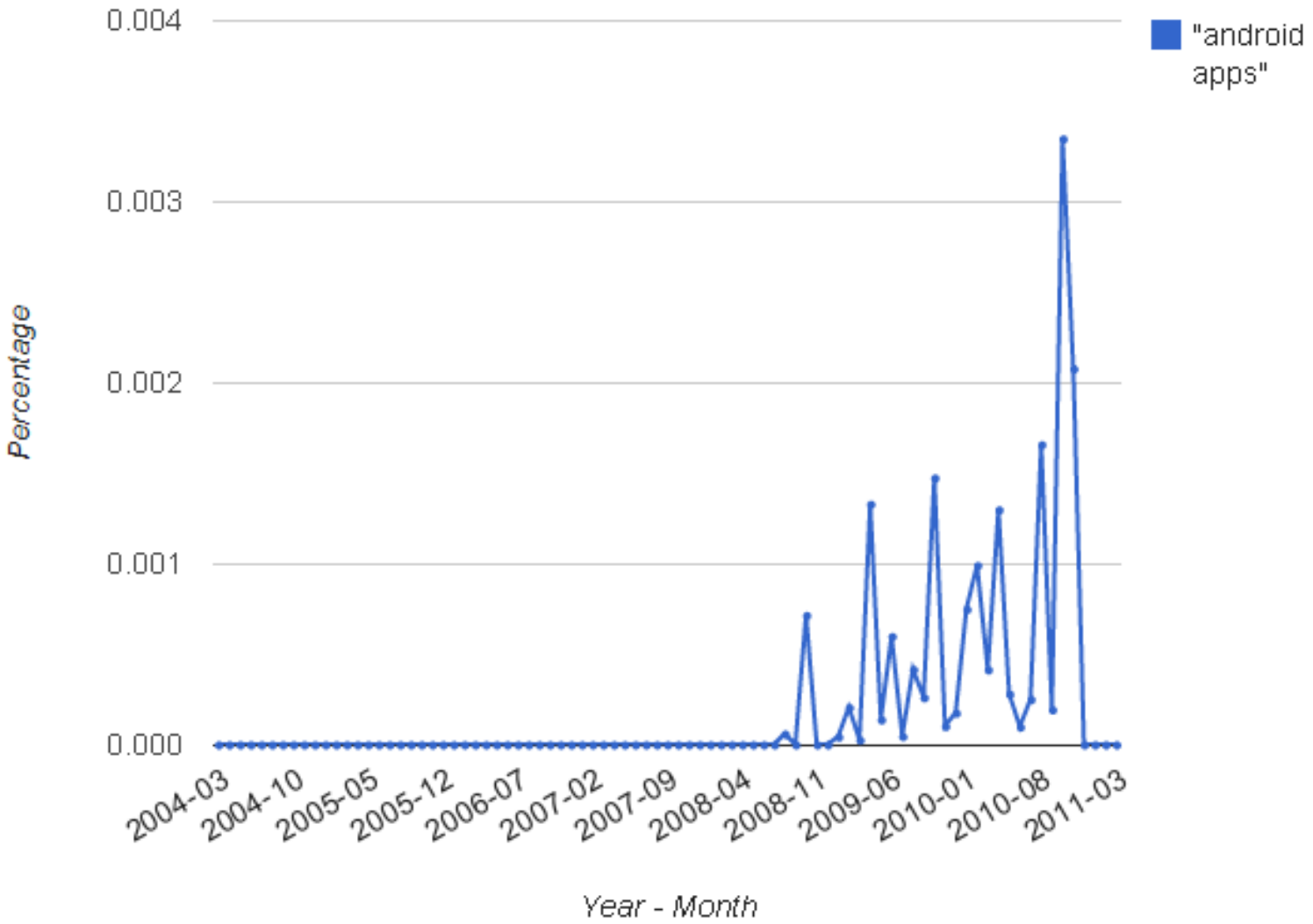
Modern access

Current interface: Ngram data visualisations

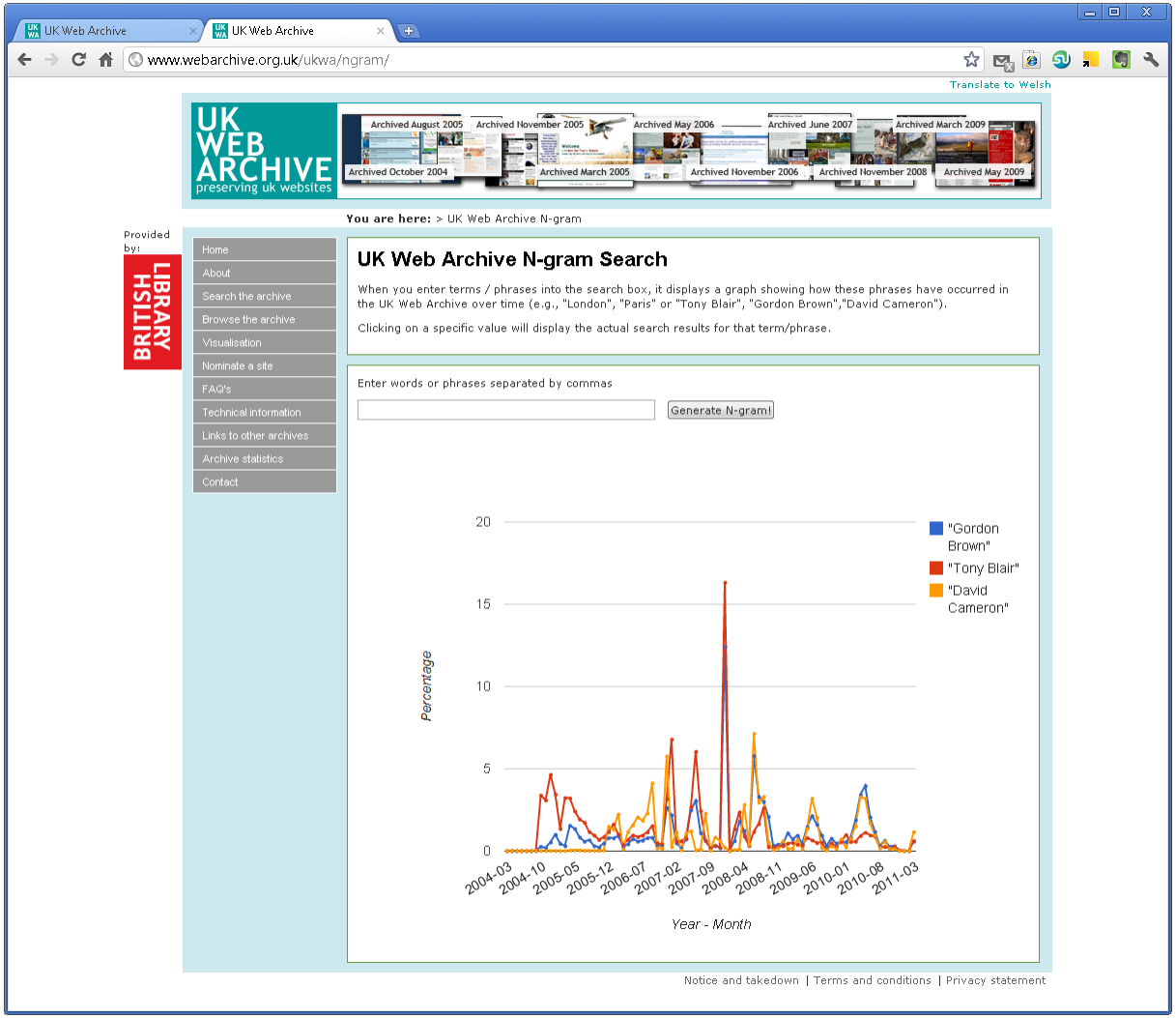


Modern access

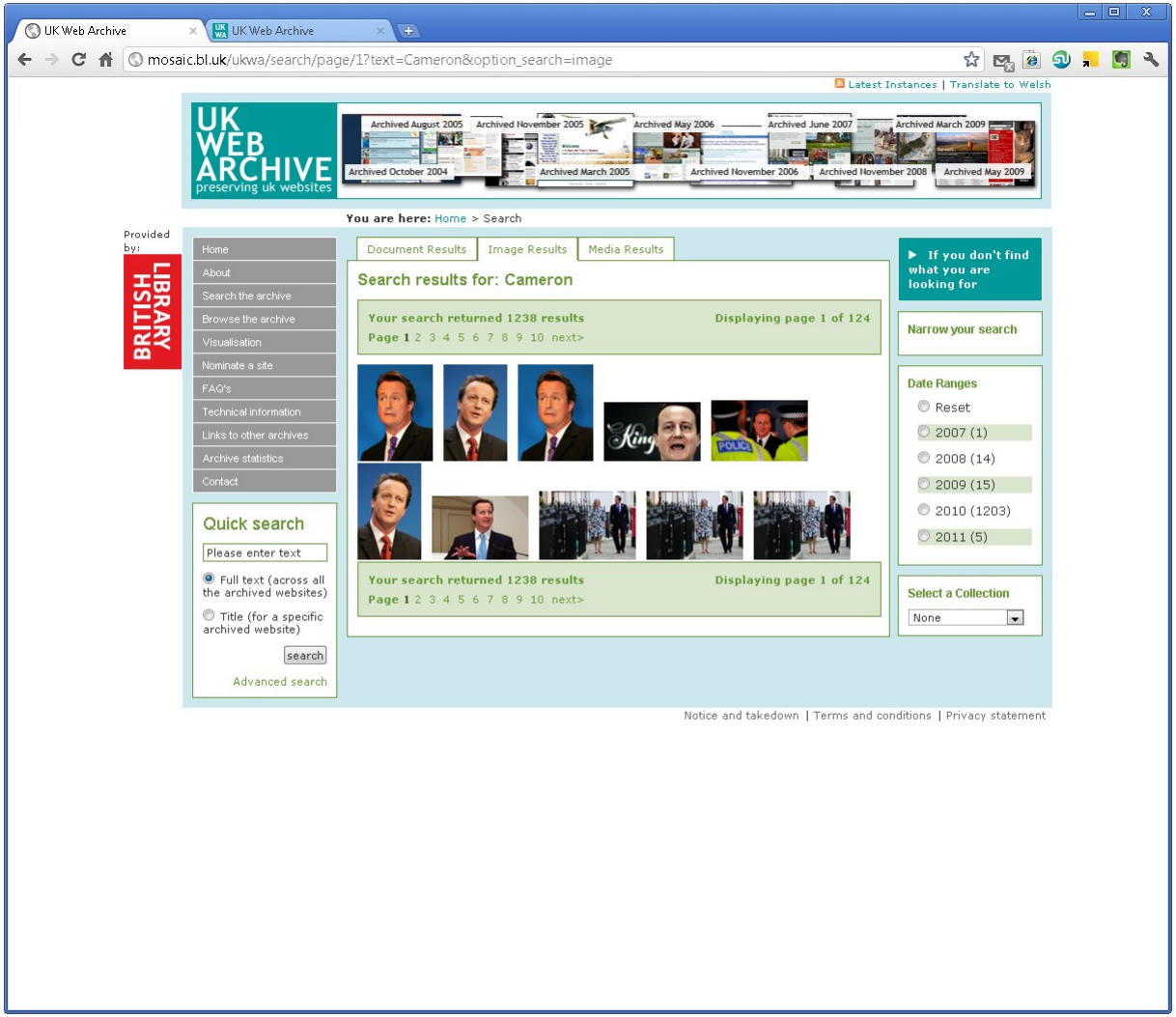
Current interface: Ngram data visualisations



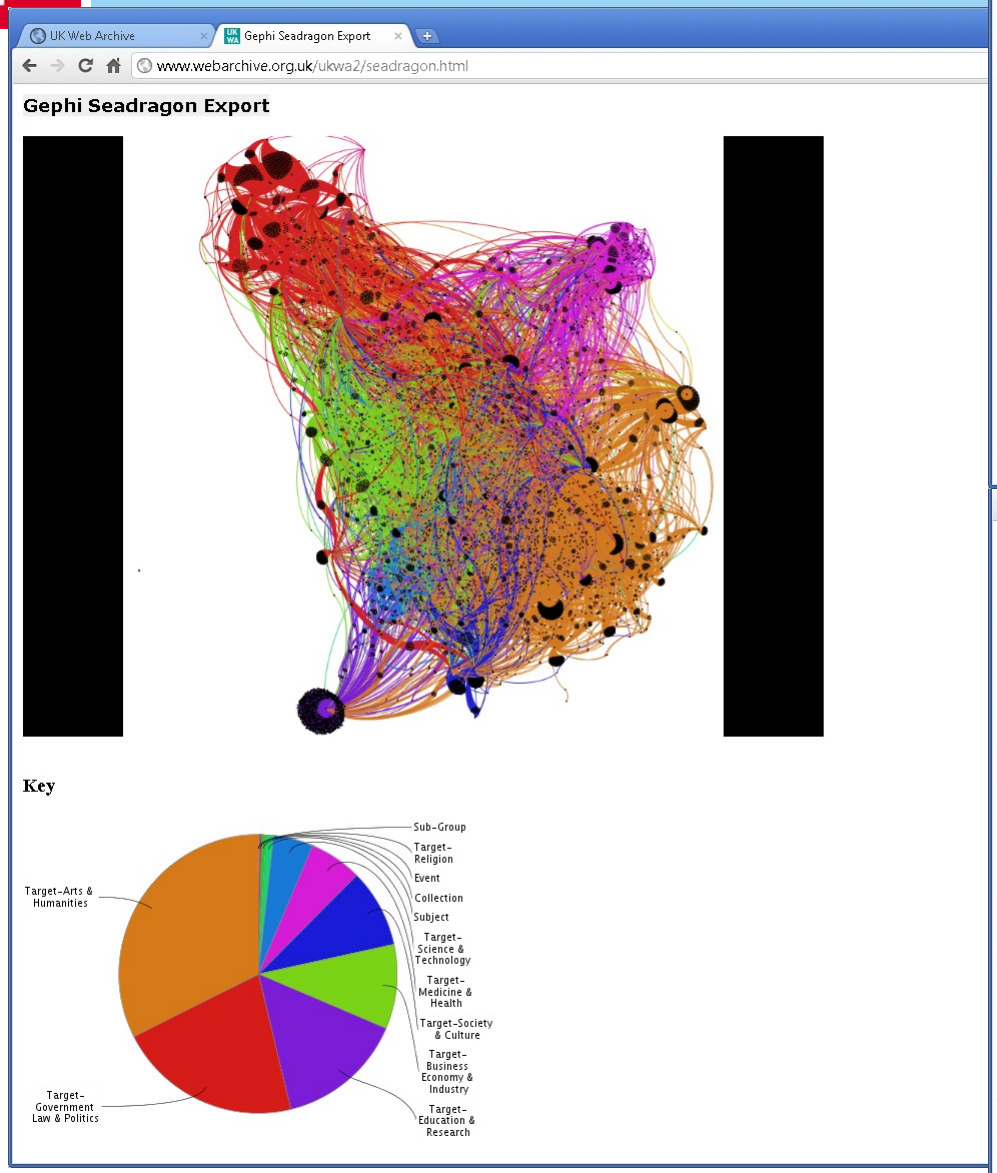
N-grams

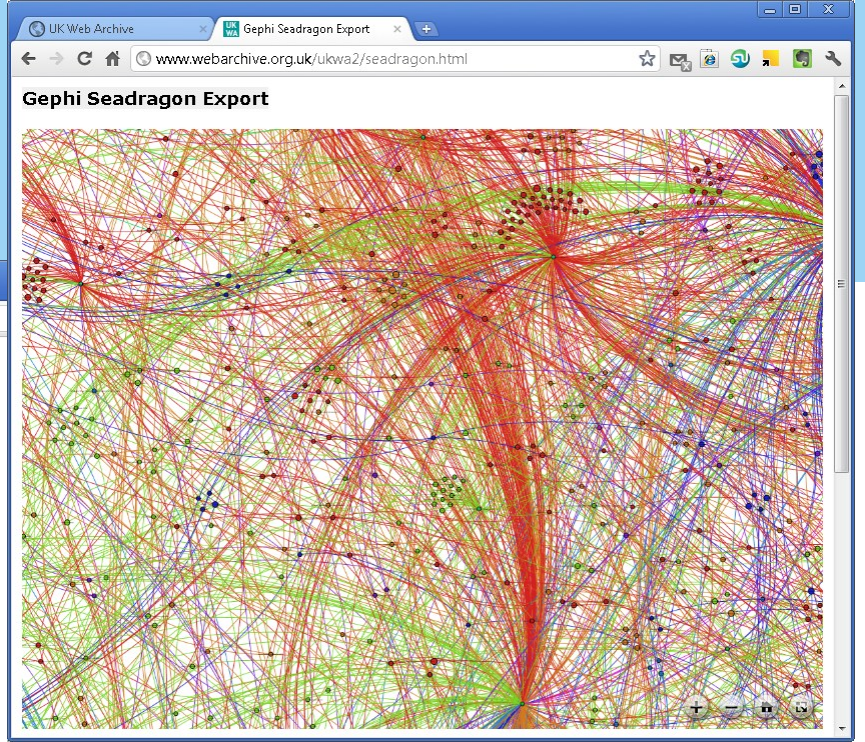

UK Web Archive
UK Web Archive
www.webarchive.org.uk/ukwa/ngram/
Translate to Welsh
UK WEB ARCHIVE
preserving uk websites
Archived August 2005
Archived November 2005
Archived May 2006
Archived June 2007
Archived March 2009
Archived October 2004
Archived March 2005
Archived November 2006
Archived November 2008
Archived May 2009
You are here: > UK Web Archive N-gram
Provided by:
BRITISH LIBRARY
Home
About
Search the archive
Browse the archive
Visualisation
Nominate a site
FAQ's
Technical information
Links to other archives
Archive statistics
Contact
UK Web Archive N-gram Search
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.
Enter words or phrases separated by commas
Generate N-gram!
20
15
10
5
0
Percentage
"Gordon Brown"
"Tony Blair"
"David Cameron"
2004-03
2004-10
2005-05
2005-12
2006-07
2007-02
2007-09
2008-04
2008-11
2009-06
2010-01
2010-08
2011-03
Year - Month
Notice and takedown | Terms and conditions | Privacy statement

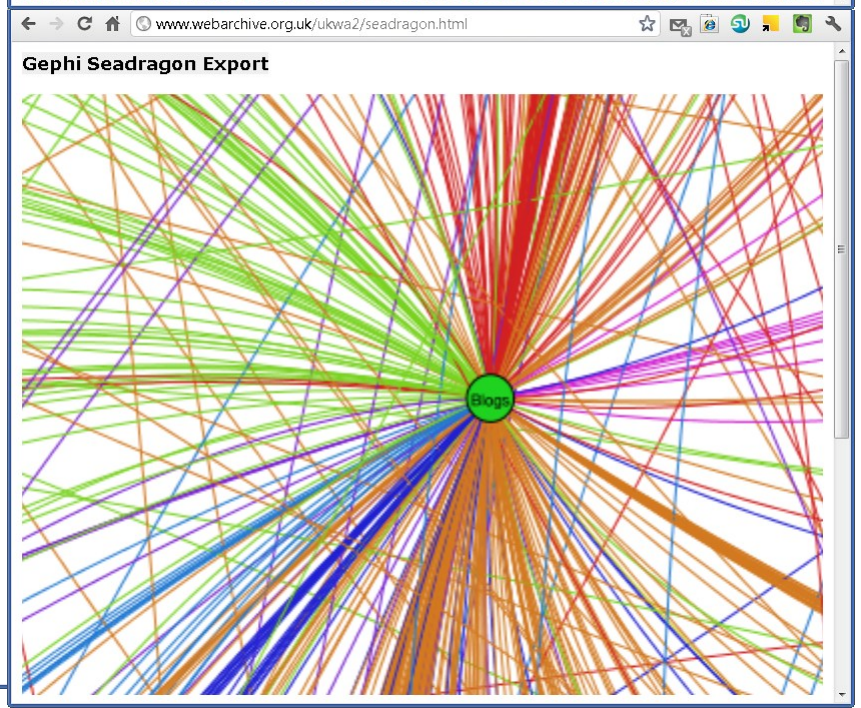
Media based Results



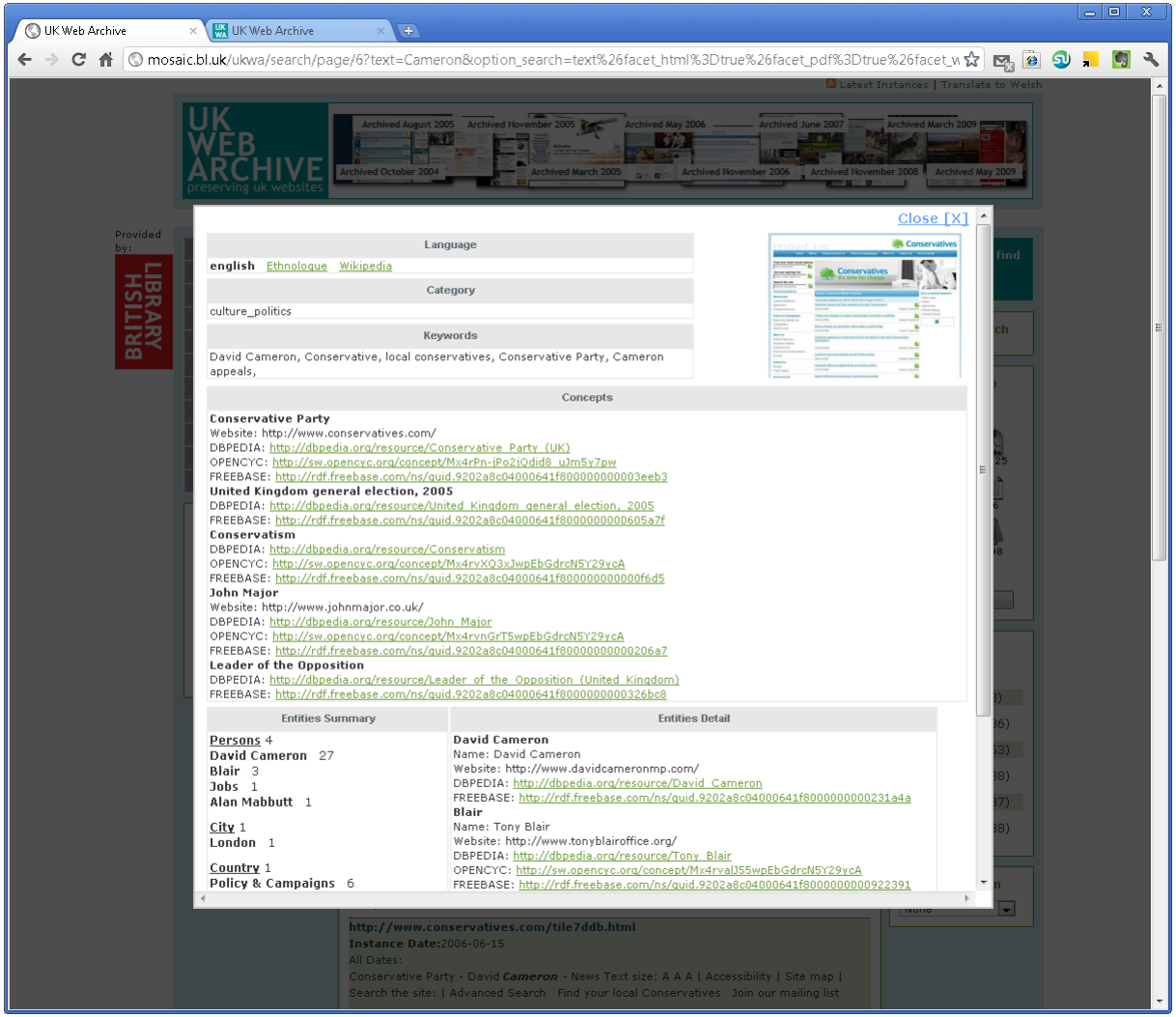
What is in the web archive?





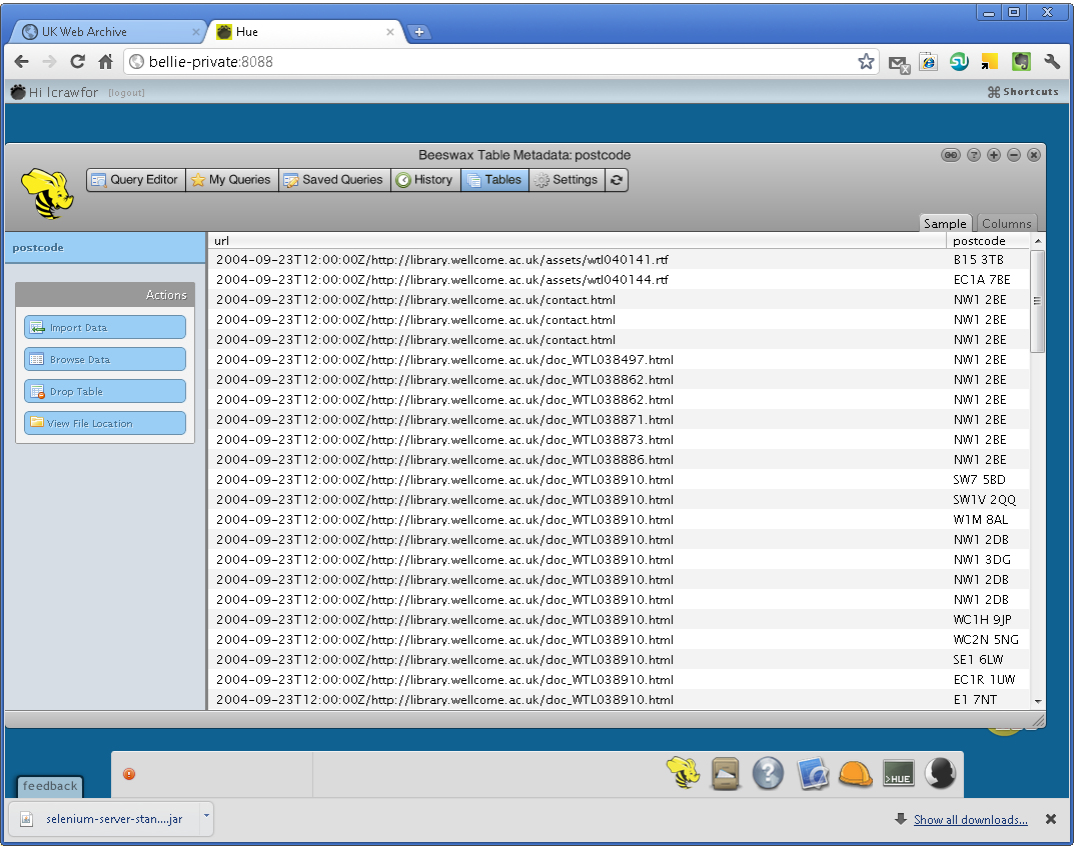


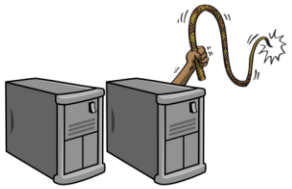
Semantic Analysis



Big Data Analytics

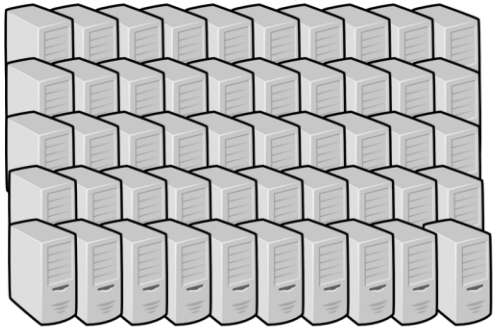
- 52 servers:
  - 1 JobTracker
  - 1 NameNode
  - 50 DataNodes/TaskTrackers



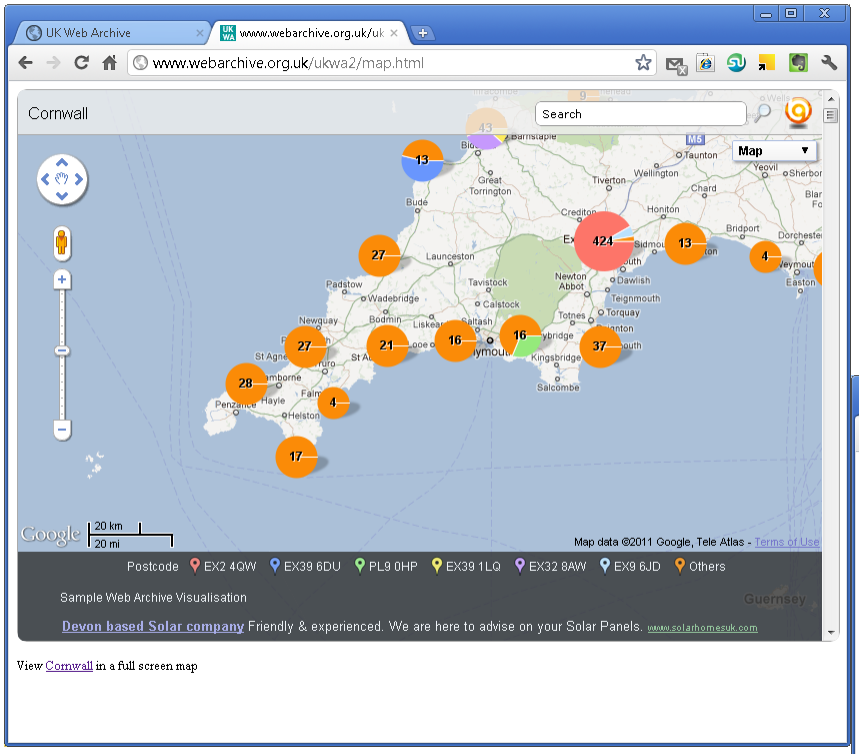


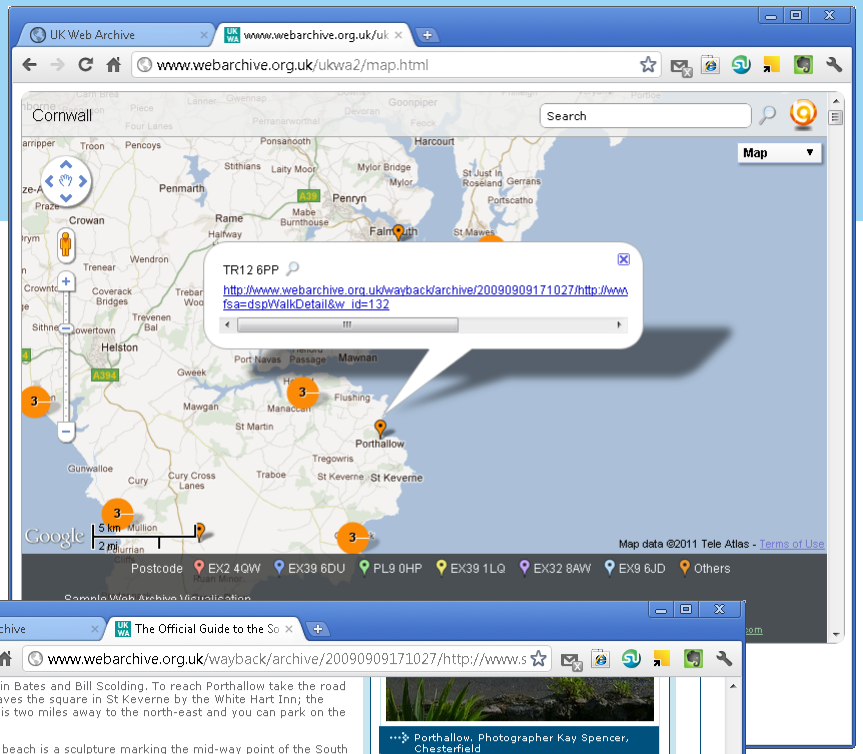
HPDL160
2 x Quad Core CPU*
64GB RAM
160GB HD

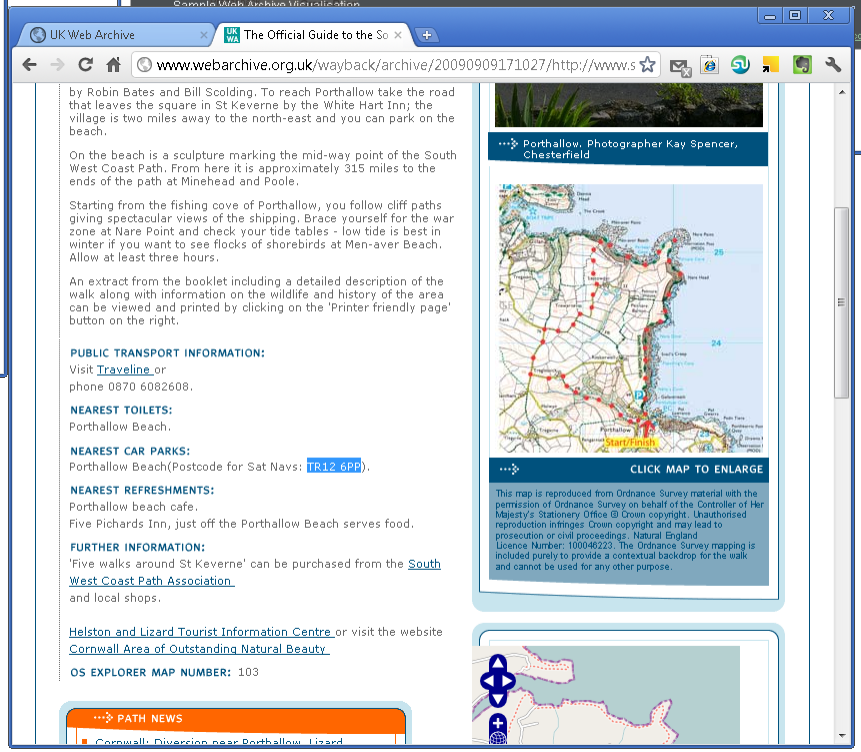
HPDL120
Quad Core CPU*
16GB RAM
60GB HD
3 x 2TB HD (HDFS)



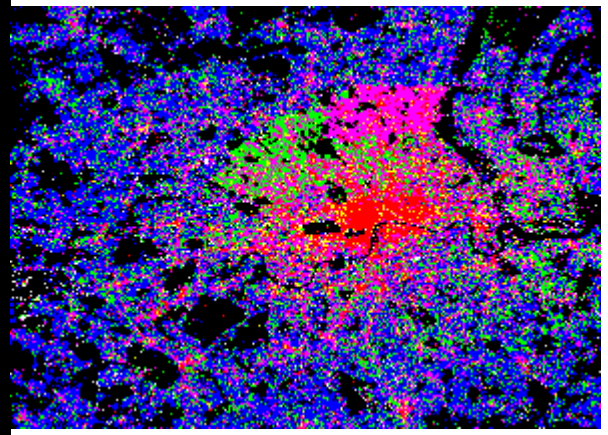
Where is the web archive?







BRITISH LIBRARY



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

BRITISH LIBRARY

Thank you!

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