Co-developing access to the UK Web Archive

Helen Hockx-Yu
Head of Web Archiving, British Library

Ten years of archiving the UK Web Archive

- Started web archiving in 2004, non-print Legal Deposit since April 2013
- Three collections: over six billion resources and over 100TB compressed data
- Focus not just on content collection
- Proactive development of access and use, through close engagement with researchers
  - User survey
  - Content selection and curation
  - Brain-storming sessions and workshops to formulate research questions
  - Research projects

JISC UK Web Domain Dataset 1996-2013

- Funded by JISC to create a research collection of historical UK websites
- Collaboration between the Internet Archive, JISC and the British Library
- Copy of subset of the Internet Archive's web collection that relates to the UK
- c.300 million resources, 60TB in total
- No local access – possible through the Internet Archive
- Can be used to generate secondary datasets

Co-design at every stage

- Research use case articulated
- Generic user requirements abstracted
- Requirements refined following feedback
- Iterative development cycles: Develop -> user testing -> feedback -> develop …

Use cases (generalised)

- Full-text/facet search -> individual resource
- Full-text/facet search -> analysis/visualisation
- Search -> corpus creation -> annotation/curation
- Corpus creation -> full-text search -> individual resource
- Corpus -> search -> analysis/visualisation
- [Derived datasets -> take-away]
- [Direct access to WARC/CDX -> take-away]

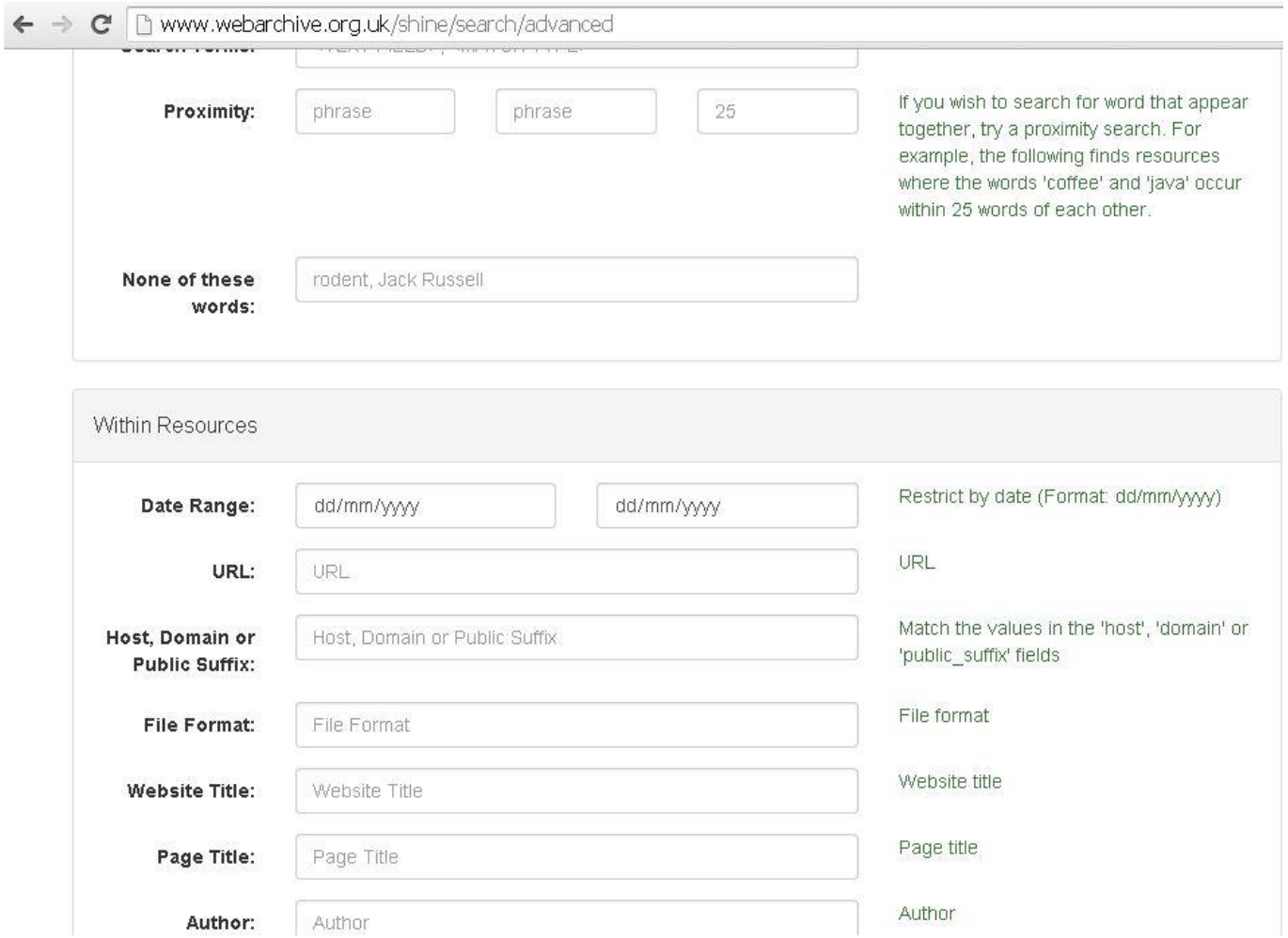
High-level requirements

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

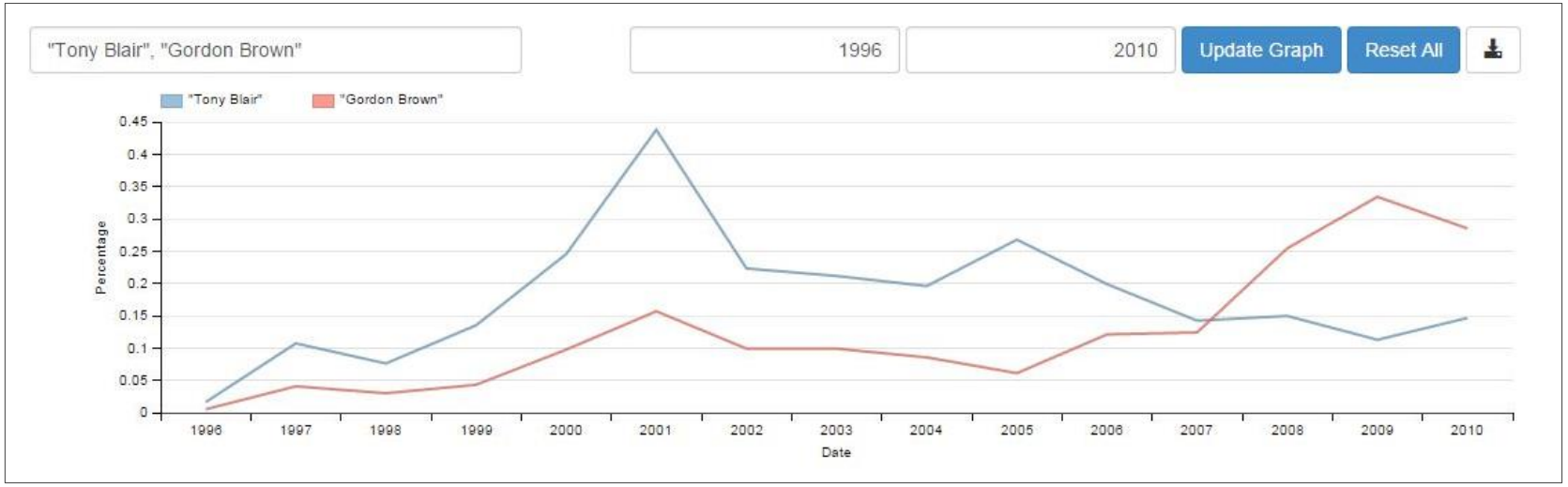
Prototype: Shine

- Full-text search, with proximity options, and to exclude specified text strings
- Apply and remove multiple facet filters to result sets
  - Content type, public suffix, domain, crawl year
  - Also available: postcode, links to public suffix, language, links domains
- Exclude single resources, or whole hosts from result sets
- Save a query
- Export basic query results, as CSV or similar
- Available at: http://webarchive.org.uk/shine

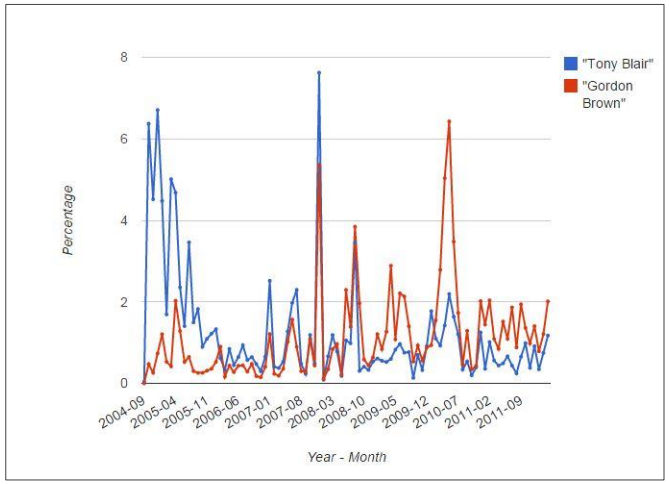
Advanced Search



Ngram



- Same search terms, different datasets
- Broadly similar trends
- Interesting to examine turning point
- Not useful without understanding of scope
- Visualisation not the end point

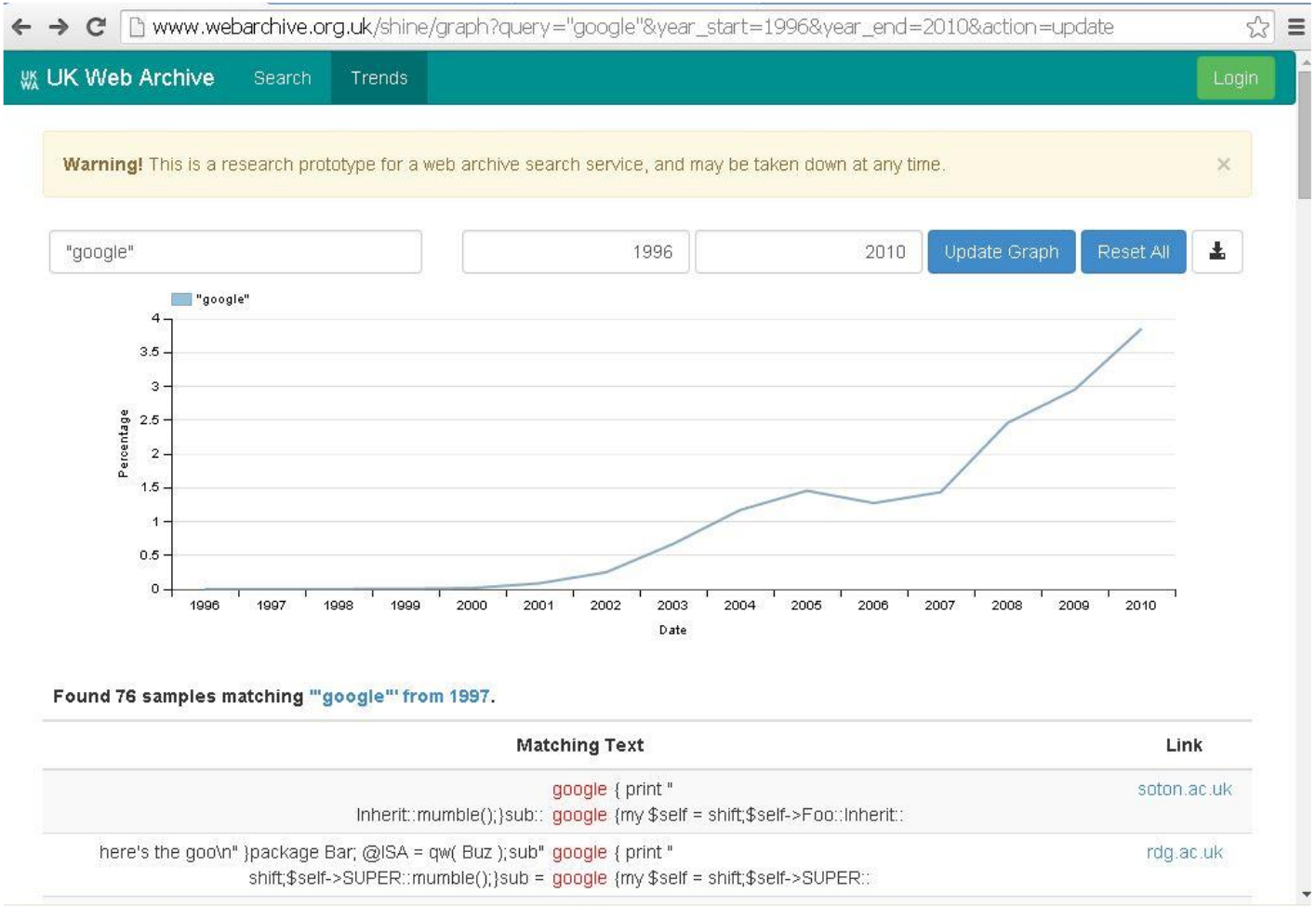


Pages mentioning “Gordon Brown” (2007)

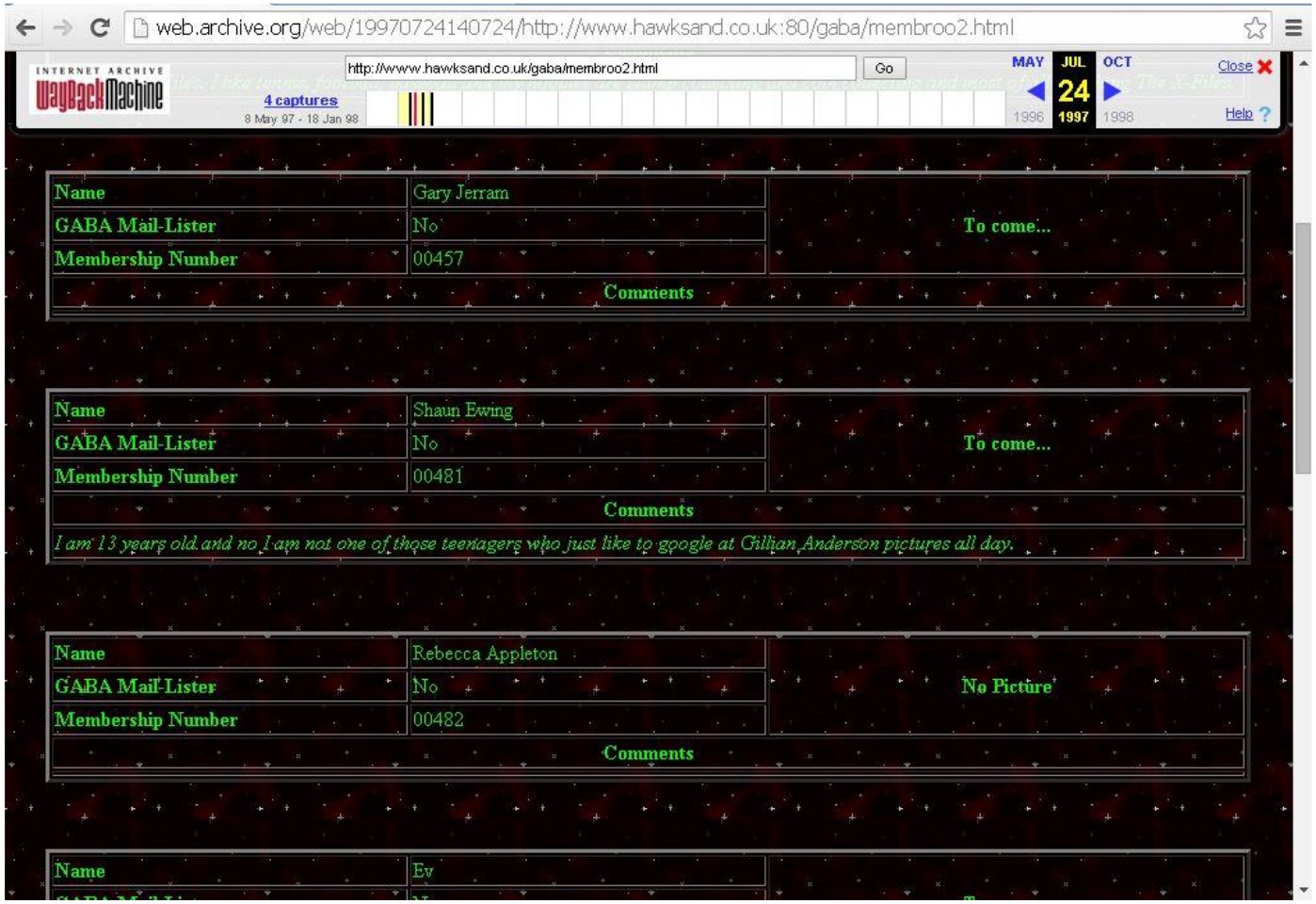
Found 100 samples matching ' "Gordon Brown"' from 2007.

	Matching Text		Link
levels before the end of 2008.So earlier this year, UK Chancellor	Gordon Brown	changed the tax regime for	bbc.co.uk
…celebrities, money, dreams, celebrity, satire, drawing, tax, bud ,	gordon brown	, lampoonEurovision Song Contest DisasterÂ Â Â Â Â Â Â Â Â Â …	humorousarts.co.uk
…Press: 1997-55HelpContact usAccess keysSite mapA-ZSearch	Gordon Brown	announced today.Mr Brown said:"The Informal ECOFIN meetin…	hm-treasury.gov.uk
and …calls to Downing Street going unanswered.advertisementHe sai	Gordon Brown Gordon Brown	, in July at the British Irish Council. Picture: PAGordon Brown w… seems to be stuck in the	theherald.co.uk
Tony Blair with all.Should those figures be taken as good news or bad news for	Gordon Brown Gordon Brown	is enough of a change to meet that public demand. Clearly the… ? Only somewhere between 30-40% of	ukpollingreport.co.uk
calling on	Gordon Brown	to build an extra 1,170 council or housing association homes in…	meltonmowbraytoday.co.uk
…BennettThursday April 19, 2007The GuardianWhoever wrote C	Gordon Brown	, he, she or they are to be	guardian.co.uk
…doing his job as Deputy PM and Minister for DETRSatisfied 32	Gordon Brown		icmresearch.co.uk
…moreDate:27/06/2007 Author:NewsBrown to be held to acc	Gordon Brown	prepares to take	theecologist.co.uk
,putative successor	Gordon Brown	, on moderation. He deplored the naive language of counter-te…	guardian.co.uk
the extent to which …according to key staff in the immigration system.Members of t ,	Gordon Brown Gordon Brown	is interested in grabbing headlines not solving problems,' Davis… this summer as a major new counter-terrorism initiative, claim t…	guardian.co.uk
…topDEPUTY ROLEHilary Benn: Wants to be deputy party leade Harman: Argues her good relationship with	Gordon Brown Gordon Brown	who is deputy - she was once shadow Chief Secretary to the	bbc.co.uk
responsible public servant should do thisAt the same time	Gordon Brown	gave this money with one hand, he was	worcesterstandard.co.uk
challenge …huge levels of fraudand error in the Tax Credits system is disgr …hints at change of heart over supercasinos - Foster.Commenti …amount of money wasted in the first three years is on course to - Huhne.Responding to	Gordon Brown Gordon Brown Gordon Brown Gordon Brown Gordon Brown	andDavid Cameron to back up their rhetoric by endorsing our p… haswalked into ’s hint during Prime walked into ’s announcement of a Â£14 million supportpackage for …	wellslibdems.org.uk
…Ministry of Defence | Defence News | Brown: UK will continue h ."Prime Minister CentreBasra handover is imminentPrime Minister	Gordon Brown Gordon Brown Gordon Brown	has reiterated the UK's position on Iraq following an announce… "I intend to make a more detailed statement when Parliament r… , making a surprise visit to troops in Iraq, has	www.mod.uk
,to persuade the chancellor	Gordon Brown	, and the prime minister to make childcare a theme for a	guardian.co.uk
…doorstep. It's time for a change of Government, as that pasty- '	Gordon Brown	is just as bad as his	thesun.co.uk
comes from no less a source than …speak fluently. However, a brief 15 minutes of catching CNN br	Gordon Brown Gordon Brown	. While the Penguinâs were away in Germany over Christmas in the	politicalpenguin.org.uk
…burmaÂ show | burmaÂ websiteGordon Brown's Lord Mayor… | …sparked severe flood warnings and evacuations on England's e	Gordon Brown Gordon Brown	will address has held	britishblogs.co.uk
gutted when speakers. Just like a lot of Labour people feel gutted when	Gordon Brown Gordon Brown	invites Maggy Thatcher to No 10. How many mosques would gi… invites Maggy Thatcher to No 10	guardian.co.uk
planning gain supplement2005 issue 48Chancellor	Gordon Brown	set to announce further consultation in pre	building.co.uk
responsible public servant should do thisAt the same time	Gordon Brown	gave this money with one hand, he was	worcesterstandard.co.uk
…Poll ResultsNewsÂ |Â Site MapÂ |Â About UsÂ - Select Country	Gordon Brown	4 67%Lifestyle Change 1 17%Total voters: 6/info/poll_results.a…	apartmentsjavea.co.uk
…ANALYSISBrown's agendaWhat sort of British prime minister ,	Gordon Brown	make?Departures loungeOutgoing British	bbc.co.uk
know the score here or that, come the 27th June, it will be	Gordon Brown Gordon Brown	weeks after the Conservative leader passed on the idea in priv… theyâll be dealing with to	ministryoftruth.org.uk

Trends analysis



Access to data supporting trends



Next steps

- Inclusion of the full JISC dataset – seamless interface to all 3 components of UK Web Archive
- Better support for corpus creation (eg combination of existing corpus)
- Annotation and sharing of corpus
- (standard) analysis and visualisation of corpus
- Faceted search within user-define corpus
- (semantic) clustering of search results

Lessons learnt

- A learning process for both
- Not a choice between “big data” or “small data”
- “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”
- Context and paratext just as important
- User expectation / assumption
- Maximum transparency
- Scale remains a challenge