Archiving the UK Web

Digital Preservation: What I Wish I Knew Before I Started
24 January 2012

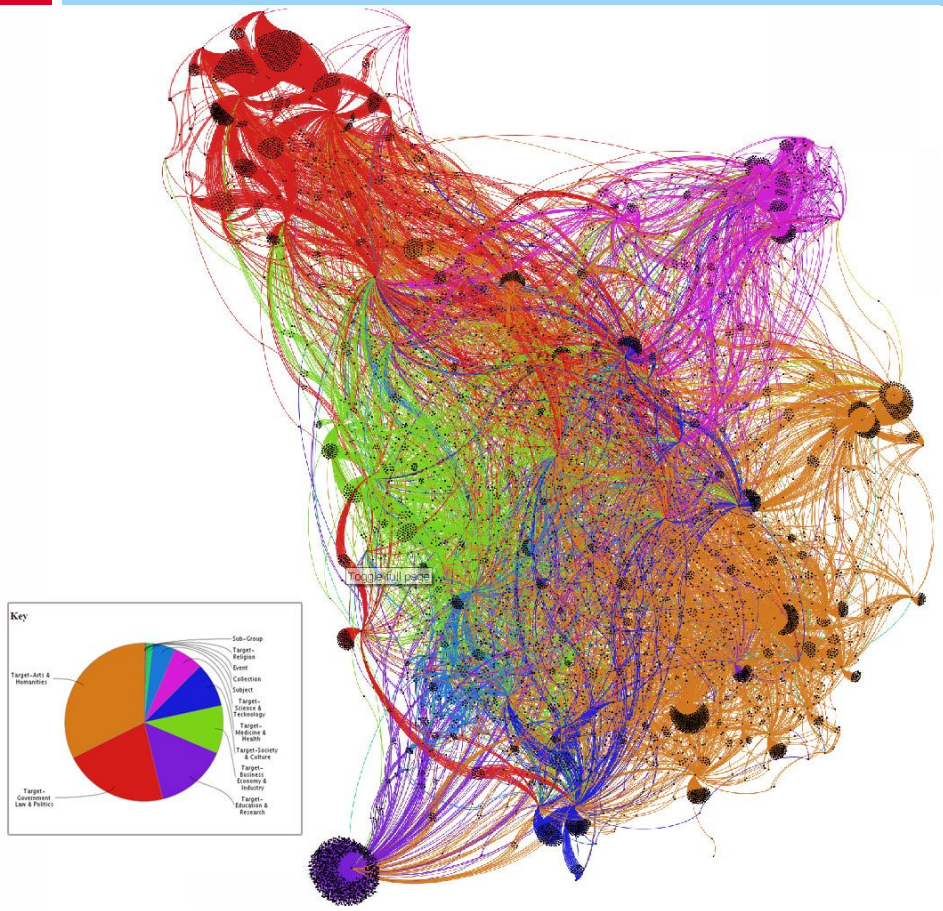
Helen Hockx-Yu

Head of Web Archiving
British Library

Web archiving: the basics

- **What**
  - Selecting, capturing, storing, preserving and managing access to snapshots of websites over time
- **How**
  - Use crawler software to download websites automatically
  - Selective or domain archiving
  - Provide access in a Web Archive
- **When**
  - Since mid 1990s
- **Who**
  - Heritage and memory organisations, eg (IIPC)
  - University libraries
  - Not-for-profit and commercial organisations, eg Internet Archive
  - Individual researchers
- **Why**
  - Global information resource
  - Artefact of cultural and technology change
  - Representative sample of the web: historical and sociological data that may not be found elsewhere
  - Part of national digital heritage - legal requirements

Scale: needle and haystack



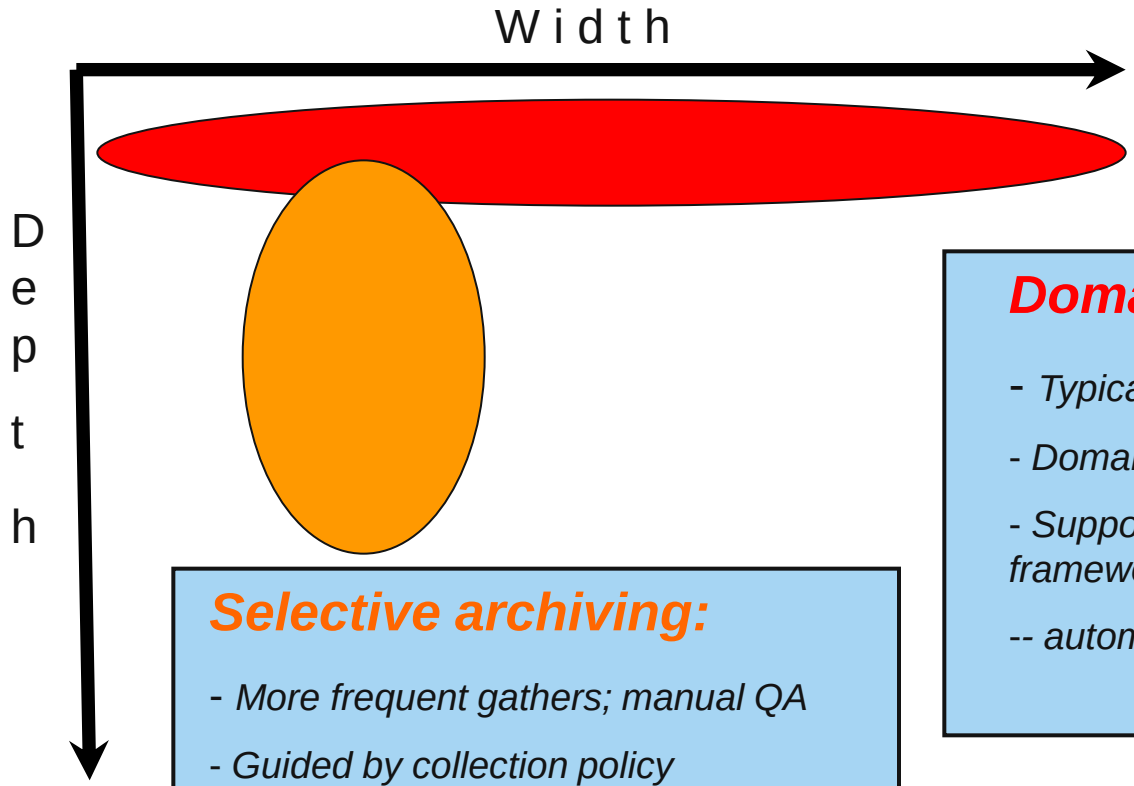
Subject hierarchy visualisation [UK Web Archive]

- over~ 10,000 websites collected since 2004
- ~ 41,000 instances

- Google: *"seen 1 trillion unique URLs"*
- more than a billion new pages are added to the web every day
- The UK web domain
  - 9 million .uk domain names registered in December 2010
  - ~ 1 million using other domain names
  - Growing at 11% - 14% per year
  - 40% estimated to be in scope for Legal Deposit
  - Estimated ~110TB each UK domain crawl

Selective versus domain archiving

- Two complementary approaches: selective and domain archiving



Selective archiving:

- *More frequent gathers; manual QA*
- *Guided by collection policy*
- *Can be based on events or themes e.g. credit crunch*
- *-- manual & expensive*

Domain harvesting:

- *Typically once/twice a year*
- *Domain wide snapshot*
- *Supported by national legislative framework*
- *-- automated & cost-effective*

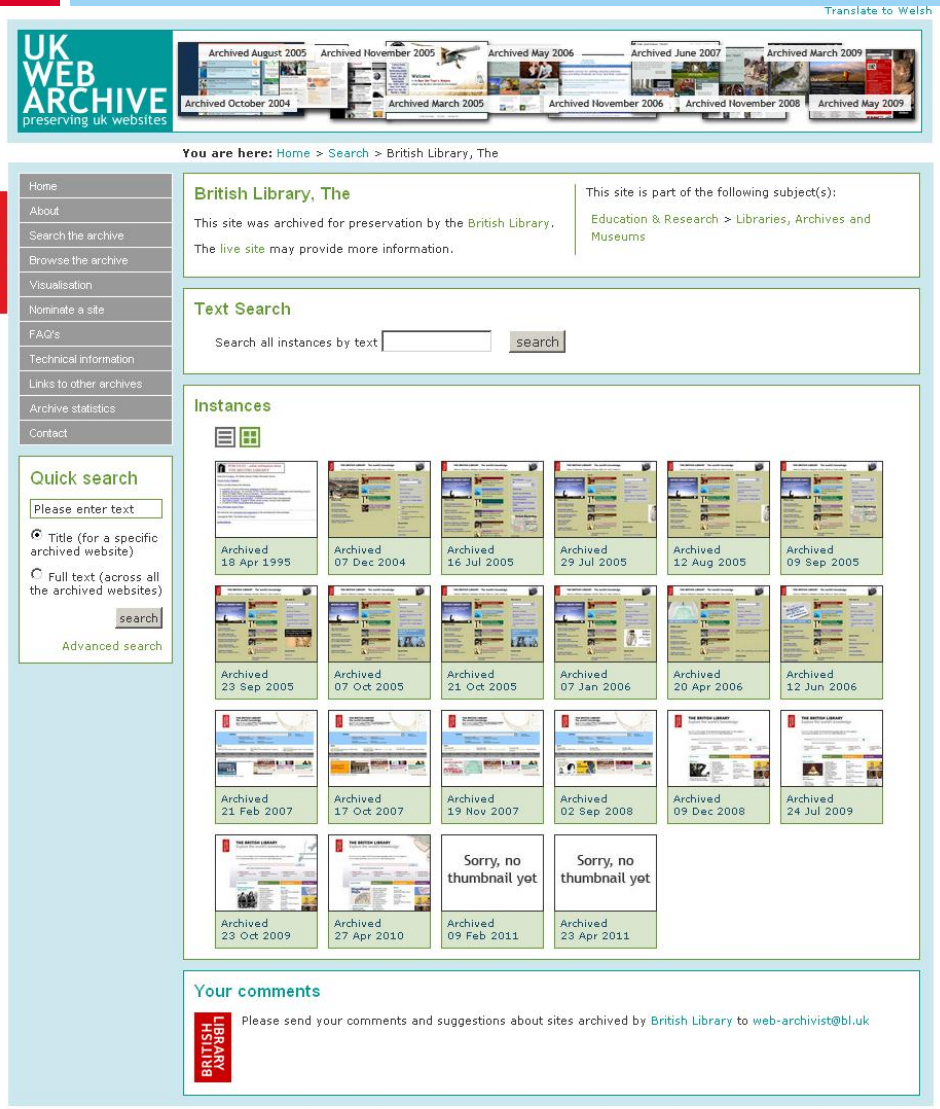
Key processes of web archiving

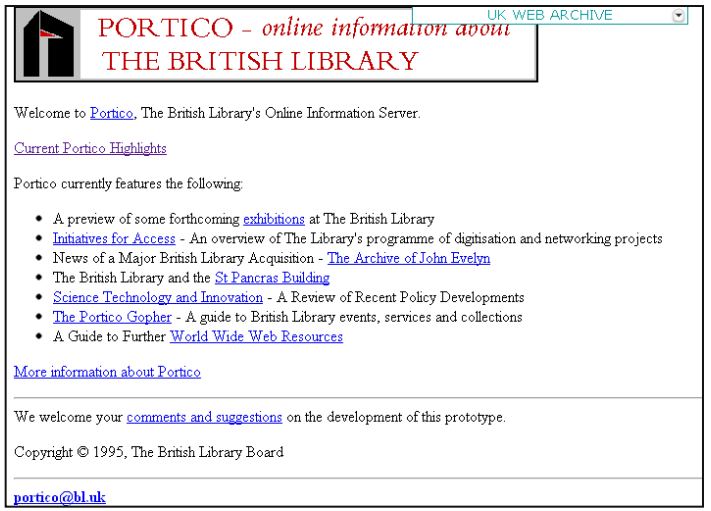
- Selection – decide what websites to archive and to include as part of a web archive collection
  - The British Library Collection Development Policy for Websites
- Automated downloading of selected websites using crawler software
- Storage
  - Archival format, eg WARC
- Access and use
  - Currently 3 ways to access the UK Web Archive http://www.webarchive.org.uk/
  - Catalogue records (at special collection level)
  - Integrated search (Primo)
- Digital preservation
  - Common and hard problem faced by all – work with the experts
  - Bit-level preservation – long term integrity of ingested bits
  - Describe the digital objects we have: metadata profile, WARC, document original (technical) environment of websites

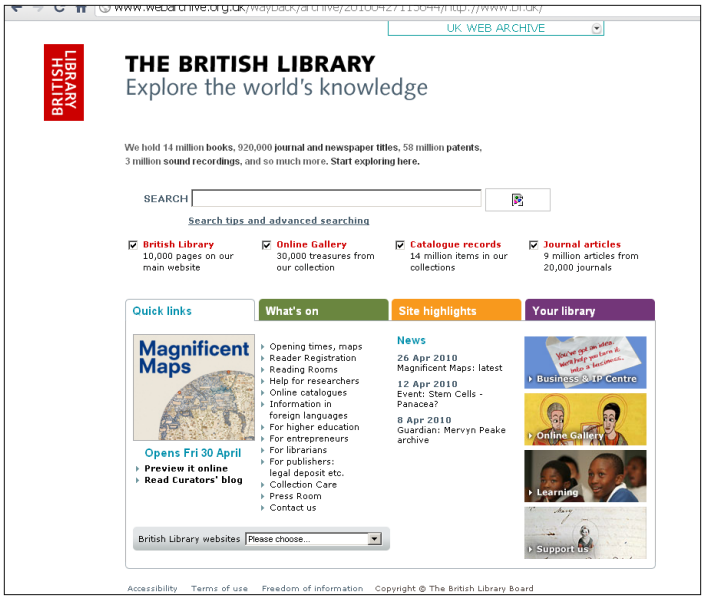
Web archiving paradoxes

- Small, closed community of practitioners – need research & reaching out to other communities
- Doubts and scepticism from various quarters
- Traditional “document-centric” approach does not scale up - canonical mission of heritage institutions being challenged
- Many technical challenges – the constant need to respond to the evolving web
  - Harvests are at best snapshots or samples
    - cannot get everything: resource and legal constraints; robot.txtexclusion, protected content
    - do not get every version: rate of change
    - the issue of temporal consistency
  - Crawler works well with HTML but struggles to capture advanced web content, e.g. rich media, dynamic and interactive content
  - “Bad” content
    - search engine spam, scam / malwaresites
    - Inadvertent ‘traps’
    - Illegal content
  - Rendering software does not always “replay” the archived content
    - Cannot reply streaming media
    - “live leakage”
- Access problem
  - Restricted access
  - Where are the users and what do they want?
- Legal issues
  - Risks of “republishing” – libel, copyright
  - Legal Deposit offers some protection but access restricted to premises of LD institutions

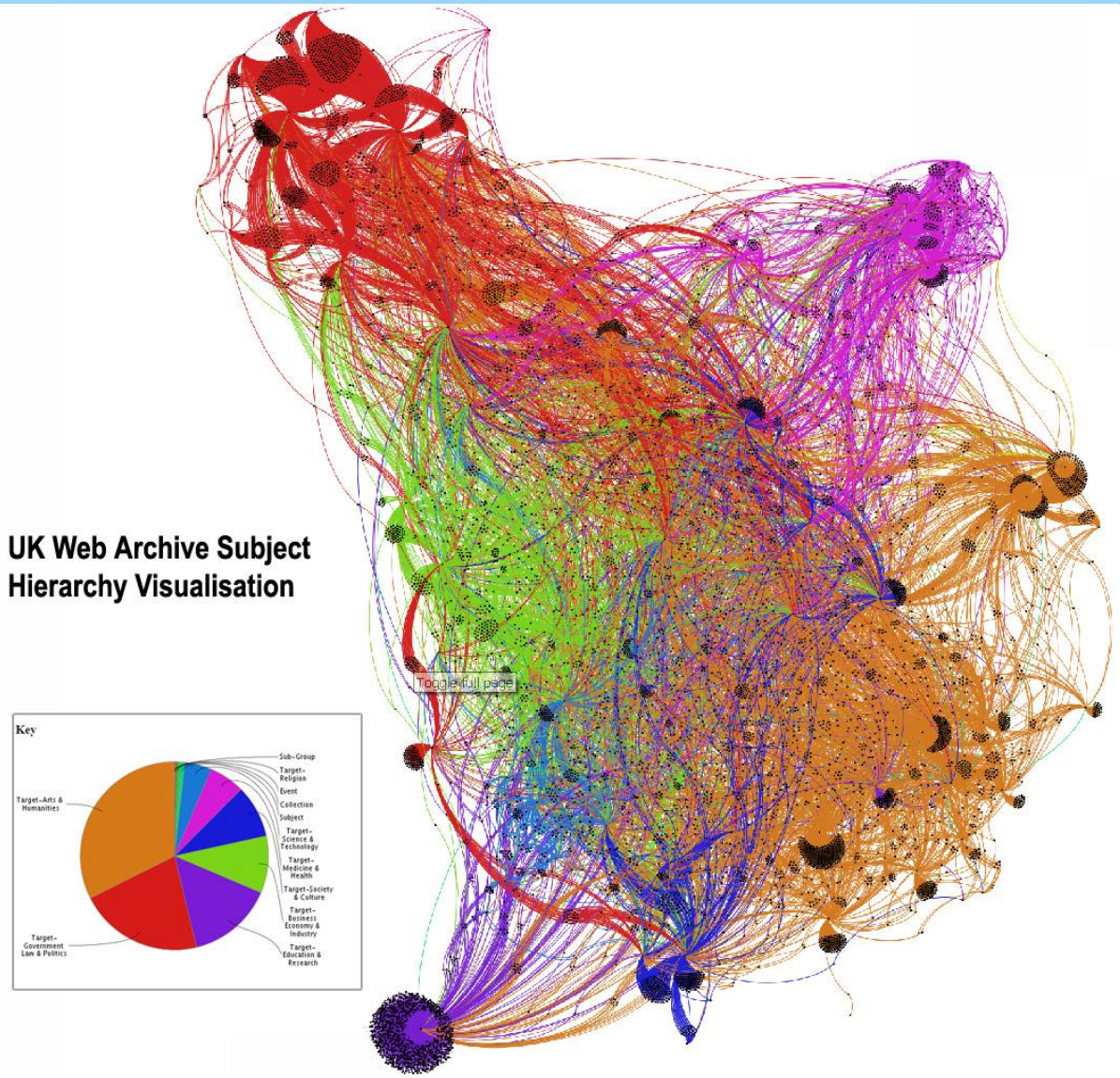
Web archive as historical documents



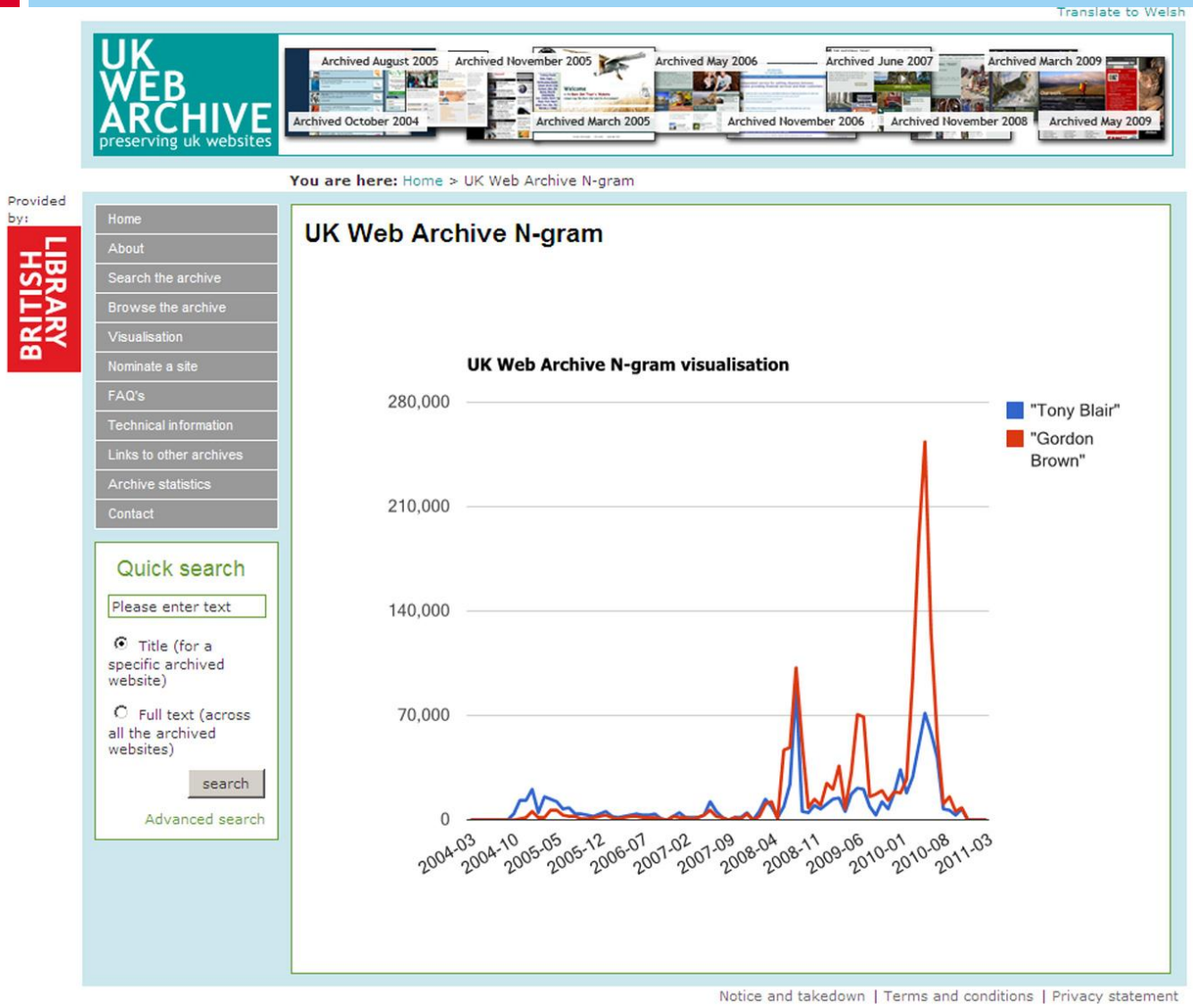




The value of the haystacks – content visualisation



The value of the haystacks - N-Gram search



Conclusion

- 14 years of web archiving – significant progress
- Yet plenty of scope for further development
- Look beyond current practices and take advantage of technologies designed for live web
- Shift of focus
  - From single page or site to entirety of web archive collection - not just for reference but also for analytics
  - Human to machine access
- Continue to grow the UK Web Archive
  - Representative of UK domain
  - used for scholarly research in a range of disciplines
  - known as the place where researchers and general public look for inactive and/or historical versions of UK websites