The Past Issue of the Web

ACM Web Science Conference 2011
14-17 June Koblenz, Germany

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

- 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

Subject hierarchy visualisation [UK Web Archive](#)
- ~10,000 websites collected since 2004
- ~37,000 instances
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.txt exclusion, 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

Welcome to Portico, The British Library's Online Information Server.

Current Portico Highlights

- A preview of some forthcoming publications of The British Library
- Initiatives for Access - An overview of The Library's programme of digitisation and networking projects
- News of a Major British Library Acquisition - The Archive of John Ruskin
- The British Library and the St Pancras Building
- Science Technology and Innovation - A review of recent policy developments
- The Portico Explorer - A guide to British Library events, services and collections
- A Guide to Further World Wide Web Resources

More information about Portico

We welcome your comments and suggestions on the development of this prototype.

Copyright © 1995, The British Library Board

Please send your comments and suggestions about sites archived by British Library to web-archive@bl.uk
The value of the haystacks – content visualisation
The value of the haystacks - N-Gram search
The value of the haystacks – postcode-based access
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
- Web Science can help!
  - Already researching many of the issues
  - Use web archives to study the web