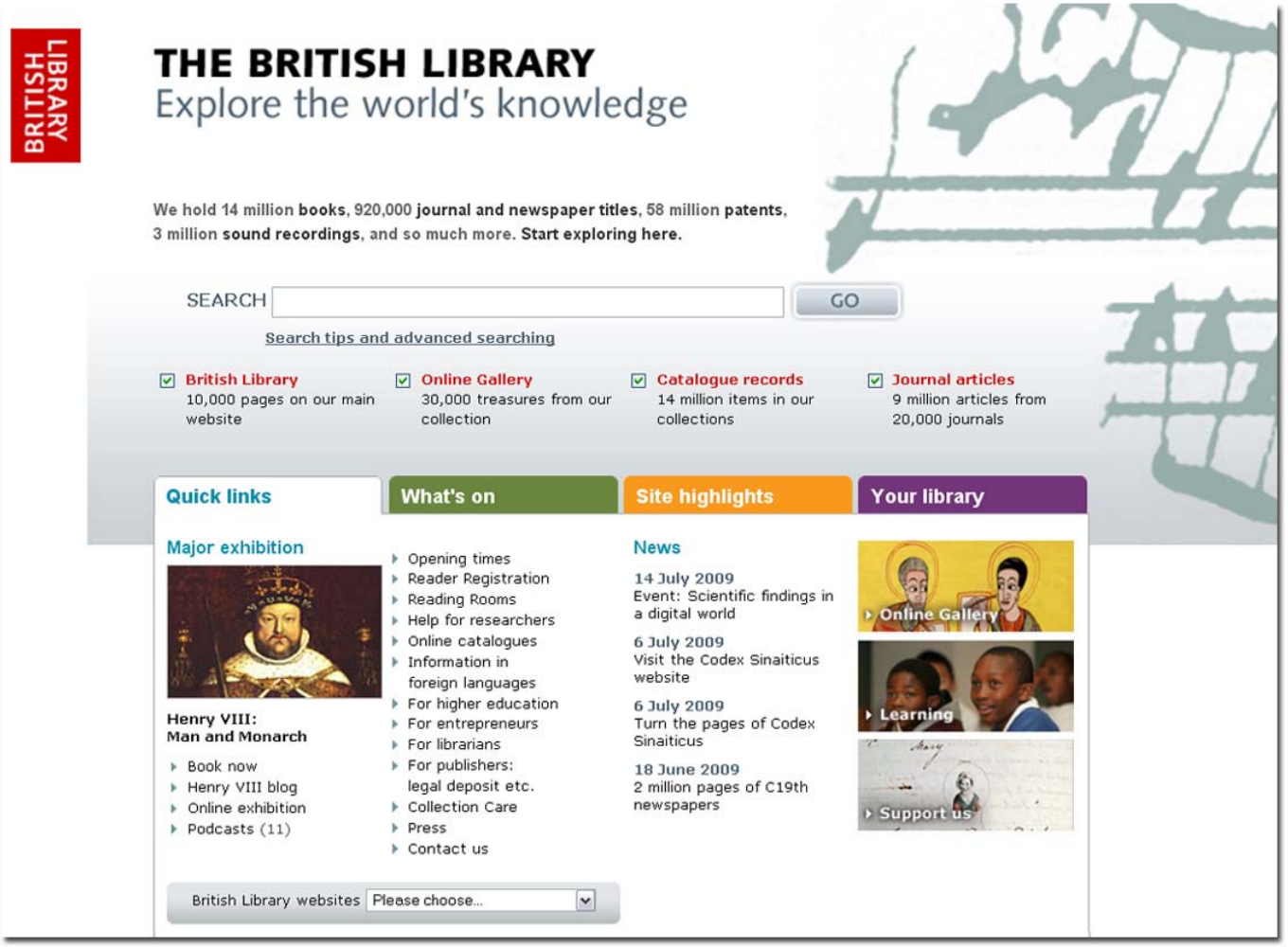
Web Archiving Tools: An Overview

JISC, the DPC and the UK Web Archiving Consortium Workshop
Missing links: the enduring web

Helen Hockx-Yu
Web Archiving Programme Manager

July 2009

Shape of the Web: HTML Pages



One “page”:

- 39 URLs or links
- 9 Images/gif
- 4 Images/jpg
- 4 JavaScript
- 4 CSS

Size of the Web

- Google: “seen *1 trillion* unique URLs”
- The UK web domain (2008) according to Netcraft Ltd
  - hosted in the UK: 9.43M
  - .uk hosted outside UK: 2.45M
  - **Total 11.88M hostnames**
- It is getting bigger - .uk TLD growing at 11% per year
- Host names does not equal actively managed content
- A large number of sites produced automatically by domain registration or hosting service companies, advertising providers or speculative domain registrants, or search-engine optimisation companies.
- HTTP/1.1 virtual hosting and load balancing technology made it possible to host a great number of active sites on a single or relatively few IP addresses.

Key Processes of Web Archiving

- Selection – decide what to capture
- Take snapshots of websites at regular intervals – harvesting or crawling
  1. Collect a page/resource (URL)
  2. Examine for references to other pages/resources
  3. Add those to the list to be collected
  4. Go back to 1
- Store the archived material on disk
  - In the original format or in a compressed archival format
  - Virus check, integrity check
- Make the archived material accessible
  - Index the files
  - Metadata
  - Render the files
  - GUI
- Ensure the archived material are accessible for long term – digital preservation

Selection

- Based on selection policies
- Fairly manual process
- Often sits outside documented workflow
- Not well supported
- Gap in tools provision

Web Harvesting Software

- Also called “web crawlers” - a computer program that browses the web in an automated manner
- Many available but different in dimensions:
  - proprietary or open-source
  - small or large scale
  - selective or broad
  - textual or all-media/archival
  - data models and formats
- Heritrix is the most commonly used web crawler by the community, created by the Internet Archive in partnership with libraries and archives worldwide
- Packaged workflow management software, integrated or coupled with a crawler; handle permissions, job scheduling, QA, descriptive metadata etc.
  - NetArchiveSuite, by the Royal Library and the State and University Library in Demark
  - Web Curator Tool, by the British Library and the National Library of New Zealand
  - PANDAS, by the National Library of Australia

Archival Formats

- ARC: developed by the Internet Archive in 1996 as a container format for archived website files; no longer maintained

- “Web ARChive” (WARC) format is now coming into mainstream use

- Developed by members of the International Internet Preservation Consortium (IIPC)

- An international standard: ISO 28500:2009, Information and documentation -- WARC file format

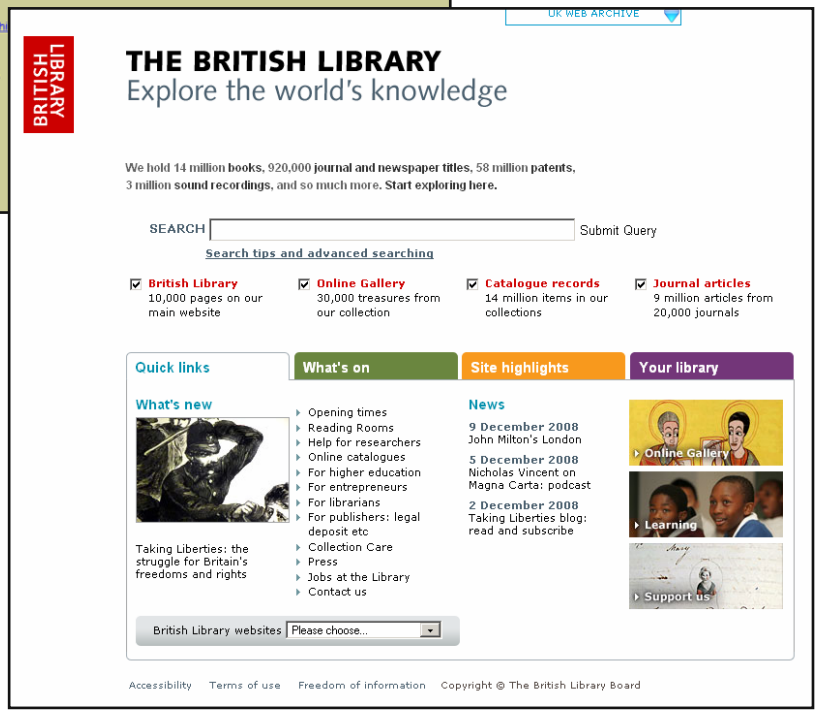
- WARC extends ARC but offers new possibilities, e.g. recording of HTTP request headers, arbitrary metadata, the allocation of identifier for every contained file, management of duplicates and migrated records, the segmentation of the records

- Designed to support long-term preservation of web archives

Accessing Web Archives

- Need a range of software
- Indexing and “replay” or rendering software, offering URL-based look-up and browsing
  - Open Source Wayback Machine (OSWM) developed by the Internet Archive
  - WARC tools being developed by Hanzo Archive
- Full-text search - Google-like word, phrase search and ranking
  - Nutch/Nutchwax by Internet Archive
  - Hanzo search tools
- Customised or extended web GUI utilising Wayback
  - e.g. UK Web Archive can be browsed by Subject, Collection and alphabetical list of website names (makes use of descriptive metadata supported by Web Curator Tool)

Temporal Navigation – showing changes over time



Digital Preservation

- Much focus has been on harvesting and providing access
- No consensus on strategy, practices and specific tools
- However
  - “the new WARC format already offers all needed information for the emulation and even a mechanism to store migrated file versions within the container
  - [WARC is] “highly desirable from a long term archival standpoint”
  - ARC to WARC migration tool (Hanzo) WARC tools
  - Global format registry
  - IIPC digital preservation working group and individual IIPC member organisations carrying out promising work
    - Preservation strategies, workflows
    - Document technical environment of the web
    - Metadata
- Building on progress made by the digital preservation community

Limitations and challenges (1)

- 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 has difficulty in capturing advanced web design. e.g. JavaScript, Flash, video, dynamic and interactive content
- “Bad” content requires intervention during harvesting
  - search engine spam, scam/malware sites
  - Inadvertent ‘traps’
  - Illegal content
- Rendering software does not always “replay” the archived sites properly
  - No control on archived content
  - Browser variance

Limitations and challenges (2)

- Interface design is a challenge when displaying large number of historical snapshots of a site
- Scalability and maturity of (full-text) indexing and search tool & ranking
- Duplicates of content – need for de-duplication
- Support for and implementation of WARC takes time – still immature
  - WARC not recognised by virus-checking software
- Community relying on open-source web archiving tools which are evolving and is burdened with the development while carrying out web archiving operations
- Rapidly developing web technology

Any Questions?

Reference:
Gordon Mohr, *Archiving the Web: the How*", presentation at the IIPC General Assembly, May 2009, Ottawa, Canada

Thank you for your attention!