Workshop B: Archiving the Web

Michael Day and Maureen Pennock
Digital Curation Centre
UKOLN, University of Bath
http://www.ukoln.ac.uk/


Workshop outline

• Session 1: The context of Web archiving - Michael Day (30 minutes)
• Session 2: Archival perspectives on Web archiving - Maureen Pennock (45 minutes)
• Coffee break (15 minutes)
• Session 3: Web archiving in practice - Michael Day (45 minutes)
• Session 4: Looking to the future - Michael Day (30 minutes)
Session 1: The context of Web archiving

Michael Day

The World Wide Web (1)

- Origins in scientific community
  - CERN (early 1990s)
  - Now part of the common 'cyberinfrastructure' of science and scholarship
  - Scientists 'increasingly reliant' on Web for supporting research activities (James Hendler, 2003)
  - Helps to promotes 'open access' principles (peer-reviewed publications, data resulting from publicly-funded research)
  - Other educational roles - e.g., e-learning
The World Wide Web (2)

• Scholarly concern with the longevity of Internet references
  – Link rot problem
  – A study of three leading peer-reviewed journals showed that 13 percent of links were inactive after 3 years (Dellavalle, et al., 2003)
  – Same trends demonstrated in biomedicine, computer science, information science, …
  – Wallace Koehler's longitudinal studies show that after seven years, just 33.8 percent of a sample of Web pages persisted at their original URL

The World Wide Web (3)

• The Web now widely used across many different communities:
  – Commerce, marketing, publishing
  – Government information (e-government)
  – Personal communication
    • e.g., 44 percent of US Internet users in a 2003 survey had contributed some kind of content to the Internet
    – "The information source of first resort for millions of readers"
    - Peter Lyman (2002)
Why preserve the Web? (1)

- Cultural importance
  - National Library of Australia noted its responsibility to develop collections of library materials, regardless of format
  - Many national libraries have now developed operational or pilot Web archives, e.g.
    - Australia, Austria, China, Czech Republic, Denmark, Finland, France, Iceland, Japan, New Zealand, Norway, Slovenia, UK, USA, etc.
    - Some have made changes to legal deposit laws to accommodate Web content

Why preserve the Web (2)

- Cultural importance
  - Internet Archive
    - not-for-profit organisation, based in San Francisco
    - Acquired Web content from Alexa Internet and its own Web crawls, provides access through the Wayback Machine (http://www.archive.org/)
    - Co-operates with memory institutions on developing special collections, e.g. Library of Congress, The National Archives (UK)
    - Part of International Internet Preservation Coalition
    - Mirror of Wayback Machine at Bibliotheca Alexandrina (Egypt)
Why preserve the Web? (3)

- Web content are records of evidence
  - National archives guidance for Web managers
  - Some collection of Web sites has started
    - The National Archives UK Government Web Archive, joint project with Internet Archive
    - US National Archives and Records Administration collected snapshot of federal agency Web sites at end of the Clinton Administration
- Scholarly interest
  - Politics (Archipol), social history (Occasio), Chinese studies (DACHS)

Why preserve the Web? (4)

- Joint approaches
  - The UK Web Archiving Consortium
    - Led by the British Library
    - Partners include The National Archives, the national libraries of Wales and Scotland, the Joint Information Systems Committee, and the Wellcome Trust
    - Sharing costs, risks and experiences
    - Each partner focuses on sites relevant to their own interests
Approaches (1)

• Automatic harvesting
  – Web crawler programs
  – National libraries tend to focus on national Web domains, e.g. Kulturaw³ (Sweden)
  – Harvester fed set of links, pages fetched, analysed, etc., etc.
  – Internet Archive uses same approach for whole Web, since 1996 has generated ~2 petabytes
    • Problems with functionality and country representation
      (but still a very valuable resource)
  – Development of Heritrix crawler program

Approaches (2)

• Selective capture or deposit
  – Pioneered by National Library of Australia (PANDORA)
  – Development of selection guidelines, selection of sites, negotiation with site owners, capture using gathering or mirroring tools
  – Used by UK Web Archiving Consortium
  – Sites can also be captured and deposited by Web site owners
    • e.g., NARA 2001
Approaches (3)

- Combined approaches
  - Some selective capture, periodic whole domain harvesting
  - Reflects relative strengths of the two approaches
    - Harvesting approach much cheaper per terabyte, enables large collections to be built up
    - More detailed attention can be paid to complex sites, e.g. database driven (deep Web) sites
  - Approach pioneered by Bibliothèque nationale de France (BnF)
  - Recent Australian whole domain harvest

Approaches (4)

- International Internet Preservation Consortium (IIPC)
  - Group of national libraries and the Internet Archive, led by BnF
  - Co-operation on coverage and access - a global distributed collection
  - Development of tools
    - Harvesting - Heritrix, DeepArc
    - Storage - ARC, BAT
    - Search and navigation - NutchWAX, WERA, Zinq
  - Web Archiving Metadata Set
Issues (1)

- What is the Web?
  - A conceptual problem
  - Components of the Web easier to understand than the whole
  - What is it that we want to preserve?
    - Content? - easy for HTML pages, more difficult for databases (or database-driven sites)
    - Interfaces?
      - Personalisation features
      - Web 2.0

Issues (2)

- Legal problems
  - Legal environment in many countries does not take Web archives into account (Charlesworth, 2003)
  - Problems with:
    - Copyright
    - Archives could be deemed to be the "publishers" of defamatory or otherwise illegal content, or held responsible for breaches of data protection legislation
  - Remedies = select content or restrict access
Issues (3)

• Scale
  – Web is large (and growing)
  – Regular snapshots grow even bigger
  – Internet Archive: almost 2 petabytes, growing at >20 terabytes a month
  – Differences in Web archive size depending on domain:
    • Finland (2002) 500 gigabytes
    • Portugal (2003) 78 gigabytes
    • Australia (2005) 6.69 terabytes

Issues (4)

• Dynamic nature of the Web
  – Pages, sites, domains, constantly changing
    • e.g. new top level domains
    • Web content disappearing (link rot)
  – Some ad hoc focus on the ephemeral
    • Political elections, sports events, 9/11, Hurricanes Katrina and Rita
  – Changes in Web technologies
    • Personalised delivery of content
    • Increased interactivity, Web 2.0, etc.
Issues (5)

- Access
  - Problem of linking content stored in multiple, distributed archives
  - Need for co-operation
  - A role for International Internet Preservation Consortium?
- Digital preservation and curation
  - What this might mean for the Web has not been explored in detail
  - Web archives need to fit into the wider landscape of digital preservation and curation initiatives

Initial conclusions

- The Web is culturally important
- To date, Web archiving initiatives have collected a significant amount of content
- Different capture techniques compliment each other
- There has been a major improvement in the tools being used to harvest and manage content, e.g. the IIPC toolkit
- Co-operation - the IIPC provides one venue for this. Are others needed?
- Many significant issues remain to be solved
Session 2:
Archival perspectives on Web archiving

Maureen Pennock
[separate presentation]

Session 3:
Web archiving in practice

Michael Day
Contexts

• National and research libraries
  – National domains
  – Special collections
• National archives
  – Snapshots of government Web sites

Selection (1)

• Develop selection policy
  – Exact criteria will depend on the purpose of the Web archive
  – National libraries will tend to focus on their role as the custodian of the nation's documentary heritage, e.g.
    • National Library of Australia
      – Selected content needs to be relevant to Australia (or written by an Australian)
      – But there is a higher degree of selectivity than in the traditional environment
      – Boundaries of document type are not so clear cut
    • Other partners in PANDORA focus on specific content
      – States, film and music, war
Selection (2)

- UK Web Archiving Consortium
  - Different member organisations focus on different content types, e.g.:
    - Medical sites (Wellcome Library), project web sites (JISC), Wales (National Library of Wales), …
  - Archives will focus on the role of Web sites as records, e.g.
    - Recording interactions between state and citizen (e-Government)

- Frequency
  - Decisions also need to be made on the frequency of capture
  - The National Archives (UK) collects some sites weekly, others biannually

Collection and ingest (1)

- Collection methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Content-driven</th>
<th>Event-driven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client-side</td>
<td>Remote harvesting</td>
<td></td>
</tr>
<tr>
<td>Server-side</td>
<td>Direct transfer Database archiving</td>
<td>Transactional archiving</td>
</tr>
</tbody>
</table>

- Source: Adrian Brown (TNA): http://www.dcc.ac.uk/events/fpw-2006/
Collection and ingest (2)

- **Direct transfer**
  - Examples:
    - NARA snapshots at the end of the Clinton Administration (2001)
    - 10 Downing Street site (2001 General Election)
  - Can be problematic, effectively a migration to a different technical environment
- **Database archiving**
  - IIPC tool developed for capture of the deep Web (DeepARC)
  - Non trivial task, mapping relational DBs into XML schema, migrating content into an XML document

Collection and ingest (3)

- **Remote harvesting**
  - The most commonly used capture method
  - Uses crawler programs similar to those used by search engines
    - To date, various crawler programs have been developed (or adapted)
    - The Internet Archive has led the development of a crawler program focused on the capture of Web content (Heritrix)
  - Collection can be focused at different levels
    - Domain capture (national domain defined in various ways), used by some national libraries
    - Focused collections, capture of selected sites
Collection and ingest (4)

- Software available to manage the capture and ingest process
  - PANDAS (Pandora Digital Archiving System)
  - For setting up crawler programs, identifying base URLs, managing harvesting parameters (for selective approach)
  - Creation of metadata
- Limitations of the harvesting approach:
  - Does not deal effectively with database-driven sites (deep Web)
  - Little quality-control of content harvested

Collection and ingest (5)

- Harvesting can also be contracted out:
  - Contracts with the Internet Archive/European Archive
    - The National Archives
      » UK Government Web Archive
      » Regular capture of selected government Web pages
      » September 11 Web Archive
      » Hurricanes Katrina and Rita Web Archive
Preservation and access (1)

- Preservation
  - Is about maintaining accessibility over time
  - About maintaining the authenticity of content (knowing that it is what it claims to be)
  - The 'significant properties' of objects are important
- Web archiving initiatives have, until now, mostly been about collecting content rather than preserving it
  - Reflects the rapidly changing nature of the Web
  - An essential first step
  - Preservation is a much harder issue to solve

Preservation and access (2)

- Preservation involves
  - The development of a secure repository system
    - e.g., based on the Reference Model for an Open Archival Information System (ISO 14721:2003)
  - Good system administration
    - Access control, management of storage (media refreshment, backup and replication), disaster recovery
  - Activities specific to digital preservation:
    - Identifying the significant properties of objects
    - Identifying and implementing appropriate preservation strategies
    - Preservation planning (dealing with future uncertainty)
Preservation and access (3)

• Access
  – Many challenges (see IIPC Use Cases)
  – Legal reasons mean that many Web archiving initiatives do not provide significant end-user access
  – Especially true for domain harvesting initiatives (national libraries)
  – However, some selective initiatives already allow access to captured content:
    • UK Web Archiving Consortium
    • The Pandora Archive
  – As does:
    • The Internet Archive …
Welcome to the European Laboratory for Particle Physics, located near Geneva, Switzerland and France. CERN is the birthplace of the World Wide Web. The WWW ecosystem provides a set of services to the physics experiments and the lab.
**Workshop: Archiving the Web, 28 September 2006**

- **Title:** FPO

**Search Results for Jan 01, 1996 - Sep 22, 2006**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>10</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

* denotes when site was updated.
Session 4: Looking to the future

Michael Day

Legal issues (1)

- General observations
  - I am not a lawyer!
  - There is much legal uncertainty in the digital domain, not least about jurisdiction
- Intellectual property
  - Copyright regimes getting more stringent (e.g., DCMA)
  - Rights holders more determined to protect IPR
  - This is the reason why the UK Web Archiving Consortium negotiates deposit of content with rights holders
    - But it can still be difficult to identify who holds the rights in multi-partner project Web sites
Legal issues (2)

- Content liability:
  - In the UK, providing access to a preserved Web site counts as "publication," raising the issue of content liability for:
    - Defamation
      - Most UK case law relates to the role of ISPs, but Web archives would seem to be liable if defamatory content is "republished"
    - Data Protection
      - Where Web pages might contain personal information, Web archives need to comply with DP legislation

Legal issues (3)

- Content liability (continued)
  - Illegal content
    - Some types of pornography, Holocaust denial
    - Wide variance internationally, but care still needs to be taken
  - If you are thinking about doing Web archiving, you will at some point need to consider legal issues, even if only to dismiss them!
Future proofing your web site (1)

- Some general principles
  - From John Kunze (California Digital Library)
  - 3 Rs
    - Reduce dependencies
    - Redirect URLs
    - Replicate
  - Prioritise
    - Focus on that content that is most important (or may contain essential business records)
  - Look for simple solutions
    - Focus on the things that may have the widest impact.

Future proofing your Web site (2)

- Basics:
  - Develop a strategy for managing Web sites over the short to medium term
    - Plan for the future, try to obtain sufficient funding
  - Maintain domain names
    - Expired names can be reused by Web site pirates
    - This can cause severe embarrassment
  - Where possible, use standards
    - Validate standards
      - Some tools exist to do this (e.g. for X/HTML)
    - Open standards are better than proprietary formats
    - Avoid browser-specific features
Future proofing your Web site (3)

- If there is no possibility of maintaining the pages yourself:
  - Record the fact that the pages are no longer being updated
  - If necessary, hand over the site to be managed by someone else
    - A role for third party hosting services? National Libraries? The UK Web Archiving Consortium?
  - This is not just a problem for organisations, personal (or hobby) sites are probably even worse off …

Conclusions

- The Web is culturally important [and also contains records]
- To date, Web archiving initiatives have collected a significant amount of content [and this is growing rapidly]
- Different capture techniques compliment each other [but significant progress has been on the development of models for selection and ingest]
- There has been a major improvement in the tools being used to harvest and manage content, e.g. the IIPC toolkit [this work continues]
- Co-operation - the IIPC provides one venue for this. Are others needed? [Web archiving is one aspect of a much wider digital preservation problem]
- Many significant issues remain to be solved
Further reading

- Adrian Brown, *Archiving Websites: a practical guide for information management professionals* (Facet, 2006)
- Andrew Charlesworth, Legal issues relating to the archiving of Internet resources … (JISC, 2003): http://www.jisc.ac.uk/uploaded_documents/archiving_legal.pdf
- UK Web Archiving Consortium: http://www.webarchive.org.uk/
- Internet Archive: http://www.archive.org/
- European Archive: http://www.europarchive.org/
- International Internet Preservation Consortium: http://netpreserve.org/

Acknowledgements

*UKOLN* is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union, and other sources. UKOLN also receives support from the University of Bath, where it is based.

http://www.ukoln.ac.uk/

The *Digital Curation Centre* is funded by the JISC and the UK Research Councils' e-Science Core Programme.

http://www.dcc.ac.uk/