Iniciativas de preservación de la Web: una visión actual

Michael Day
Digital Curation Centre,
UKOLN, University of Bath, UK
m.day@ukoln.ac.uk

Archivo de la Internet española: Webs y archivos personales, Madrid, Spain, 12 December 2005

http://www.ukoln.ac.uk/  http://www.dcc.ac.uk/  http://www.bath.ac.uk/
Presentation overview

• Reasons for collecting and preserving the Web

• Main approaches to collection:
  – Whole-domain harvesting
  – Selective capture or deposit
  – Combined approaches
  – International Internet Preservation Consortium (IIPC)

• Issues:
  – Conceptual, legal, technical (size and dynamic nature), preservation and curation
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
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)
About the Wayback Machine

Browse through 40 billion web pages archived from 1996 to a few months ago. To start surfing the Wayback, type in the web address of a site or page where you would like to start, and press enter. Then select from the archived dates available. The resulting pages point to other archived pages at as close a date as possible. Keyword searching is not currently supported.

http://archive.bibalex.org, the Internet archive at the New Library of Alexandria, Egypt, mirrors the Wayback Machine. Try your search there when you have trouble connecting to the Wayback servers.

Wayback Machine Hardware

Web Collaborations with the Smithsonian and the Library of

The Wayback Machine

Put the Wayback Machine right in your browser!

The Wayback Machine Bookmarklet

Drag this link to your browser’s toolbar: Wayback

When you visit a page that you want to find an old version of, just click the toolbar link. You will be transported to any historic versions at the Wayback Machine.

Thanks to oxford.com

Web Collections

National Archives

The UK Central Government Web Archive is a collective collection of UK Government websites, archived from August 2003, which has been collected by the Internet Archive on behalf of the National Archives of the United Kingdom, history.
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. Kulturarw³ (Sweden)
  – Harvester fed set of links, pages fetched, analysed, etc., etc.
  – Internet Archive uses same approach for whole Web, since 1996 has generated >1 petabyte
  • 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
    - Interfaces?
  - Personalisation features
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: >1 petabyte, 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
  – Role for IIPC?

• 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
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?
- Some significant issues remain to be solved
## Search Results for Jan 01, 1996 - Dec 09, 2005

<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>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>0 pages</td>
<td>0 pages</td>
<td>2 pages</td>
<td>3 pages</td>
<td>4 pages</td>
<td>17 pages</td>
<td>7 pages</td>
<td>13 pages</td>
<td>20 pages</td>
<td>0 pages</td>
</tr>
</tbody>
</table>

**Dec 06, 1996** *Dec 12, 1996*

**Jan 25, 1999** *Apr 19, 1999*

**Mar 02, 2000**

**Oct 18, 2000**

**Nov 09, 2000**

**Dec 03, 2000** *

**Feb 04, 2001**

**Mar 01, 2001** *

**Apr 01, 2001** *

**Apr 13, 2001** *

**May 15, 2001** *

**May 23, 2001** *

**Jul 21, 2001** *

**Sep 26, 2001** *

**Nov 16, 2001** *

**Nov 18, 2001** *

**Nov 24, 2001** *

**Nov 26, 2001** *

**Nov 27, 2001** *

**Dec 06, 2001** *

**Dec 14, 2001** *

**Dec 16, 2001** *

**Dec 18, 2001** *

**Jan 24, 2002** *

**Jan 29, 2003** *

**Feb 26, 2004** *

**Apr 26, 2004** *

**May 19, 2004** *

**Jun 06, 2004** *

**Jun 09, 2004** *

**Jun 11, 2004** *

**Jun 22, 2004** *

**Jun 24, 2004** *

**Jun 26, 2004** *

**Jun 27, 2004** *

**Jul 01, 2004** *

**Jul 04, 2004** *

**Jul 20, 2004** *

**Aug 13, 2004** *

**Sep 19, 2004** *

**Oct 27, 2004** *

**Nov 19, 2004** *

**Nov 24, 2004** *

**Nov 26, 2004** *

**Nov 28, 2004** *
Thank you / gracias
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/

JISC  MLA

The Digital Curation Centre is funded by the JISC and the UK e-Science Programme: http://www.dcc.ac.uk/