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

UKOLN is supported by

JISC MLA MUSEUMS LIBRARIES ARCHIVES

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
  - "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 Fransciso
    - 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)

Edit View Go Bookmarks Tools Help

http://www.archive.org/web/web.php

Google | ISI Web of Knowled... | Journals | Catalogues | Bellringing | Dorset | Die Bahn - Homepage | Welcome to the NDIIPP

ARCHIVE

Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA

INTERNET ARCHIVE WayBackMachine

Home | Wayback Machine | Researcher Access | FreeCache | SFLan | Petabox | Open Source Media | Heritrix | BookMobile

Search: Wayback Machine GO! Advanced Search

Anonymous User (login or join us)

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

INTERNET ARCHIVE WayBackMachine

http://

Take Me Back

Advanced Search

Take The Wayback Machine With You

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 gyford.com

Web Collections

National Archives
The UK Central Government Web Archive is a selective 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. Kulturarw[3] (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 is 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

Edit View Go Bookmarks Tools Help

http://www.archive.org/ Go

ARCHIVE

Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA

Forums | FAQs | Contributions | Jobs | Donate

Search: All Media Types GO!

Universal access to human knowledge

Anonymous User (login or join us)

Announcements (more)

Dead update

Katrina web archive launches, over 25 million pages, text searchable

Web-archive-on-demand service for libraries launched

Web

40 billion pages

WaybackMachine

http://

Take Me Back Advanced Search

Welcome to the Archive

RSS

The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.

Moving Images

22,972 movies

Browse (by keyword)

Upload your own movie

This Just In (more) RSS

Technolotics #20 -...
20 minutes ago

Curator's Choice (more)

Brothers and Sisters - filmstrip

Live Music Archive

27,731 concerts

Browse (by band)

Upload your own concert

This Just In (more) RSS

The LeeVees Live at...
4 hours ago

Curator's Choice (more)

Live Music Archive

John Brown's Body Live at State Theater on...

Disc 1 1. Band Introduction 2. Singers & Players 3. Among

Audio

49,451 recordings

Browse (by keyword)

Upload your own recording

This Just In (more) RSS

wimp ma pimp
1 hour ago

Curator's Choice (more)

Presidential Recordings

Richard Nixon Press Conference 3/24/1972

Nixon press conference on March 24, 1972. The press

Texts

24,729 texts

Browse (by keyword)

This Just In (more) RSS

Seven Myths About...
9 minutes ago

Curator's Choice (more)

Edit View Go Bookmarks Tools Help

http://web.archive.org/web/*/http://www.residencia.csic.es/ Go

INTERNET ARCHIVE
WayBackMachine

Enter Web Address: http:// All Take Me Back Adv. Search Compare Archive Pages

earched for http://www.residencia.csic.es/ 67 Results

te some duplicates are not shown. See all.
denotes when site was updated.

Search Results for Jan 01, 1996 - Dec 09, 2005

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
pages	0 pages	2 pages	3 pages	4 pages	17 pages	7 pages	13 pages	20 pages	0 pages
		Dec 06, 1998 *	Jan 25, 1999	Mar 02, 2000	Feb 04, 2001	Jan 24, 2002	Jan 29, 2003	Feb 26, 2004	
		Dec 12, 1998	Apr 21, 1999 *	Oct 18, 2000	Mar 01, 2001 *	May 26, 2002 *	Feb 12, 2003	Apr 28, 2004	
			Apr 30, 1999	Nov 09, 2000	Apr 01, 2001 *	May 29, 2002 *	Mar 24, 2003	May 19, 2004	
				Dec 03, 2000 *	Apr 18, 2001	Jul 21, 2002	Jun 03, 2003	Jun 06, 2004	
					May 15, 2001 *	Sep 30, 2002	Jun 13, 2003	Jun 09, 2004	
					May 23, 2001	Nov 23, 2002 *	Jun 22, 2003	Jun 11, 2004	
					Jul 21, 2001	Nov 26, 2002	Jul 30, 2003	Jun 22, 2004	
					Sep 26, 2001		Aug 06, 2003	Jun 24, 2004	
					Nov 16, 2001		Sep 23, 2003	Jun 25, 2004	
					Nov 18, 2001		Oct 07, 2003	Jun 26, 2004	
					Nov 24, 2001		Oct 19, 2003	Jun 27, 2004	
					Nov 26, 2001		Nov 27, 2003	Jul 01, 2004	
					Nov 27, 2001		Dec 08, 2003	Jul 04, 2004	
					Dec 06, 2001			Jul 20, 2004	
					Dec 14, 2001			Aug 13, 2004	
					Dec 16, 2001			Sep 19, 2004	
					Dec 18, 2001			Oct 27, 2004	
								Nov 24, 2004	
								Nov 26, 2004	
								Nov 28, 2004	

Edit View Go Bookmarks Tools Help

http://web.archive.org/web/19990128065527/www.residencia.csic.es/pres/presenta.htm

Go

Residencia de Estudiantes

EXPOSICIÓN

UN SIGLO DE CIENCIA EN ESPAÑA

A partir del 24 de diciembre

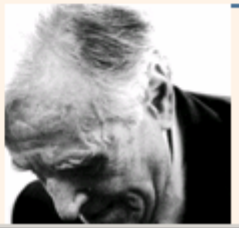
Historia de la Residencia
La Residencia Hoy
Amigos

Calendario Semanal

Calendario semanal de actos

Publicaciones
Boletín Residencia
Centro de Documentación

EXPOSICIONES



Correo
English

[From December 1998]

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/

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