WWW: The Whys and Whats of Web Archiving

Maureen Pennock
Digital Curation Centre
UKOLN, University of Bath
Overview

• What exactly is the World Wide Web?
• Why bother archiving websites?
• What you should think about about
• What you should know about
• What you could do next…
What is the World Wide Web?
Number of websites (1990 - 2008)

Source: Royal Pingdom
World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists, Policy, No W3 news, Frequently Asked Questions.

What’s out there?
Pointers to the world’s online information, subjects, W3 servers, etc.

Help
on the browser you are using

Software Products
A list of W3 project components and their current state. (e.g. Line Mode, X11 Viola, NeXTStep, Servers, Tools, Mail robot, Library)

Technical
Details of protocols, formats, program internals etc

Bibliography
Paper documentation on W3 and references.

People
A list of some people involved in the project.

History
A summary of the history of the project.

How can I help?
If you would like to support the web.

Getting code
Getting the code by anonymous FTP, etc.
The Information Superhighway is dead, long live Web 2.0?

Source: Kaidohmaru, Flickr
• What's on it now?
• More like what's *not* on it!

*Source: Sebastien Prooth, Flickr [Foobar poster from Fooboy]*
maxman (D.-Calif.), the new chairman of the House Oversight and
Investigation Committee. Waxman and others are miffed because CPCs
are withholding funds from abortion providers. He is also the Democrats’
attorney general from sexual relations. Yet going after CPCs is unlikely to
work, he said, because the issue of abortion has been losing them votes,
and moderate on the issue and even recruited them with considerable success. Now that they have taken
in the fight against abortion while not doing anything to death base, and at the same time
people on grays.” These are American voters who
are susceptible to appeals from either side.
emist both the NARAL, Barbara Boxer
ation activists who believe every single case of Roe v. Wade and political realities,
the time being, so pro-abortion forces
ach could not only divide and demonize each other but also divide and demonize pro-lifers,
the number of abortion and watch pro-lifers,

It looks like you're posting bullshit to the Internet.

What would you like to do?

- Actually check your facts for once
- Take a basic course in logical reasoning
- Enter the reality-based community
Why archive websites? (i)

• Web sites can contain uniquely available informative records and data
  • Users may act or take decisions based on this information, with important consequences
• Records of business transactions
• Accountability & transparency
  • To funding bodies
  • To stakeholders
  • For legal reasons
• Some examples...
Communicate

Ask the PM

Gordon Brown wants to hear about the issues that matter to you through a new YouTube initiative. Submit your video question and view the PM's answers in this regular feature.

Go to the Number 10 YouTube channel

...from the PM

Hear directly from the PM on specific issues of importance to the nation.

In addition to our regular newsletter...
Wiltshire County Council

Latest news articles

- **School admissions for 2009**
  Wiltshire parents are being encouraged to apply for the school they would like their children to attend next September. The online application form ...

- **Wiltshire's new council takes to the road**
  Wiltshire’s new council is taking to the road to get local people involved in its future. Wiltshire Council will come into effect in April 2009, when ...

More News...

Featured Items

- Can't find what you're looking for...
- Towards One Council
- Equality Action Plan Consultation
- Learn about the Beyond the Immediate Project

What's On

- The Wiltshire Open at Salisbury Library
- Warminster Library Readers Group
- Noisy Storytime at Salisbury Library
- Harry Potter Team Read Special at Trowbridge Library
- Read Wiltshire Rath lain
Welcome to the University of Bath

SoA Annual Conference :: York :: August 2008

Students: Welcome to Bath!

News and events

- Action imminent for Team GB’s University of Bath based modern pentathlon contingent
- Researchers study how cancer cells ‘come unstuck’ & produce secondaries
- Student designs prize-winning Olympic starting blocks
- University of Bath launches MSc in Innovation & Technology Management

more news
podcasts and videos | video intro

University of Bath, Bath, BA2 7AY, UK · tel 01225 388388
© 2007 · disclaimer · privacy statement · FoI · updated: 27 August, 2008 by the web team
ATTENTION: All Users!

Virus Warning**

Users are reminded not to open e-mails from unknown senders which contain attachments, particularly if the subject line refers to "Statement of Fees 2008/09" or similar. For more information, see the guidance from ICS at http://www.dundee.ac.uk/ics/services/virus/

System Upgrade

My Dundee has now been upgraded to version 8.0.260.7. Thank you to the ICS and Learning Centre teams for achieving this.

- Please report any problems or glitches to vle@dundee.ac.uk in the usual way.

- Please log any problems with the ICS Service Desk and

Terms of Use

Using "My Dundee" means you are subject to, and agree to abide by, the following:

Code of Conduct and Regulations for the Use of Computing Facilities

Login Here

Enter login information here and click the Login button below.

Username: 
Password: 

Login

Change Your Password

Forgotten or need to reset your password? Visit the ICS Passw Portal. Remember this will change your password for the netgroup Groupwise and My Dundee. More password information and adv available from ICS. (Close the new browser window to return to Dundee - if you have reset your password you may need to loc
Why archive websites? (ii)

• Cultural heritage objects: reflection of modern society
Why archive websites? (ii)

- Cultural heritage objects: reflection of modern society
Why archive websites? (ii)

- Cultural heritage objects: reflection of modern society
Social networks: linking people on a global scale

Source: Luc Legay, Flickr
Questions to ask yourself (i)

• What exactly do I want to archive?
  • My own website?
    • Just the back-end data?
    • Records hosted on the website?
    • Metadata about the website?
    • User generated website data?
  • Other people’s websites?

• Why do I want to archive them?
Questions to ask yourself (ii)

• Who owns the websites and their content?
• Do I have legal permission to collect and re-host the content?
• Which websites do I want to archive?
• Where will the money for this come from?
• How will I do all this?
  • Should I do it myself or rely on others?
• Who else if doing it and what tools do they have?
Good job there’s an Internet Archive then!

Source: Mr Wright, Flickr
But…
So what’s the technical problem?
So what’s the technical problem?
UKWAC

• UK Web Archiving Consortium (6 members)
  • British Library, National Library of Scotland, National Library of Wales, The National Archives, Wellcome Library, JISC
• Collects Web content selectively according to individual interests
  • Uses modified collection/harvesting software developed by the National Library of Australia
  • Permission is sought from site owners in advance
  • Allocates Persistent Identifier URLs
  • Partners assumes responsibility for their ‘own’ sites
  • Central repository of metadata
  • The collections are publicly accessible
• Website: http://www.webarchive.org.uk/
IIPC

- International Internet Preservation Consortium
  - Co-ordinated by the National and University Library of Iceland
  - UK National Archives also a member
- Online toolkit that addresses:
  - Acquisition
  - Focussed selection and verification
  - Collection storage & maintenance
  - Access & finding aids
- Website: http://www.netpreserve.org
Web Continuity project

• Led by the UK National Archives
• Comprehensive of government websites by TNA
• Redirection tool to direct users to archived content if original is no longer available
• Guidance for government webmasters on best practice for design and maintenance
• Due to complete in November 2008
• See http://www.nationalarchives.gov.uk/webcontinuity/
What you could do next (i)

• Identify your requirements: what do I want to archive and why?
• Seek help from the experts
• Develop a written policy and strategy to support activities and help secure resources
• Take a life-cycle approach to support curation and preservation planning
• Be aware of other people’s web archiving activities: check other heritage collections before gathering!
What you could do next (ii)

- Plan for preservation activities to maintain access to authentic resources over time and avoid incurring extra costs
- Determine access and user requirements costs
- Re-assess your strategy on a very regular basis
Thank You

Maureen Pennock
m.pennock@ukoln.ac.uk
http://www.dcc.ac.uk