



Web Archiving

Frank McCown
Introduction to Web Science
Harding University
Spring 2011

What is it?

- **Web archiving** is the process of collecting pages from the Web and saving them in an archive
- Usually it's important to save all associated resources (images, style sheets, etc.) to preserve look
- Archives are typically produced using web crawlers

The Ephemeral Web

- **Link rot** is a significant problem
 - Kahle ('97) - Average page lifetime is 44 days
 - Koehler ('99, '04) - 67% URLs lost in 4 years
 - Lawrence et al. ('01) - 23%-53% URLs in CiteSeer papers invalid over 5 year span (3% of invalid URLs “unfindable”)
 - Spinellis ('03) - 27% URLs in CACM/Computer papers gone in 5 years
 - Ntoulas et al. ('04) – predicted only 20% of pages today will be accessible in a year
- Even if links don't disappear, existing content is likely to change over time

Why archive the Web?

- If the Web isn't saved, we might lose a significant amount of our digital heritage
- The Web gives historians and other social scientists significant insight into our society, especially into how technology has affected it
- Serves as important resource in many lawsuits
- Some organizations want to or are legally obliged to archive their web materials
- Someone worked hard on this stuff, why not save it?

Solo archiving

The screenshot shows a web browser window with the address bar displaying `thenextweb.com/uk/2011/02/10/172-doomed-bbc-websites-saved-by-one-geek-for-3-99`. The page layout includes a navigation sidebar on the left with categories like 'Navigation', 'ALL STORIES', 'TOP STORIES', 'CHANNELS', 'APPS', 'APPLE', 'DAILY DOSE', 'DESIGN & DEV', 'ENTREPRENEUR', 'EVENTS', 'FACEBOOK', 'GADGETS', 'GOOGLE', 'INDUSTRY', 'LIFEHACKS', and 'LOCATION'. The main content area features the article title '172 doomed BBC websites saved by one geek, for \$3.99' and the author '10/02/2011, Martin Bryant'. Below the title are social sharing buttons for Twitter (253), Facebook (155), and a search bar containing '172 doomed BBC website'. An 'Update at foot of post.' section follows, containing a paragraph of text and a photograph of the BBC logo.

Navigation

PREV IN UK 10/02/2011, Martin Bryant NEXT IN UK

172 doomed BBC websites saved by one geek, for \$3.99

Tweet 253 f SHARE 155 172 doomed BBC website

Update at foot of post.

As we recently [reported](#), the BBC is set to close down 200 of its websites in the near future as part of cost-cutting measures. Hearing that 172 of these sites would be deleted from the Web entirely, an anonymous (and possibly even "Anonymous", judging by the picture on the website) individual has [taken matters into his or her own hands](#).



If you could choose a website
to preserve for all time,
what would you choose?

What websites will people want to
look at 50 to 500 years from now?

K12 Web Archiving Program

- Program by Internet Archive and Library of Congress
- 5th to 12th graders get to participate in web archiving activities
- As of Oct 2010, students in the program have archived 2,379 websites

Video:

<http://www.archive.org/details/K12WebArchivingProgramStudents>



Who is archiving the Web?

- Internet Archive



- Founded by Brewster Kahle in 1996
- Largest web archive in the world (150+ billion pages)
- Pages Available to public via Wayback Machine
- Also collects old recordings, books, video, and other digital works

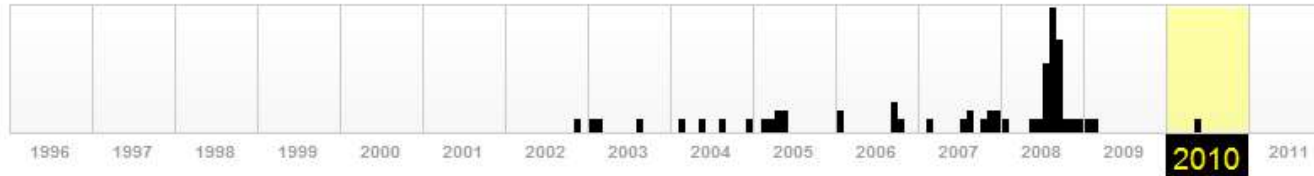




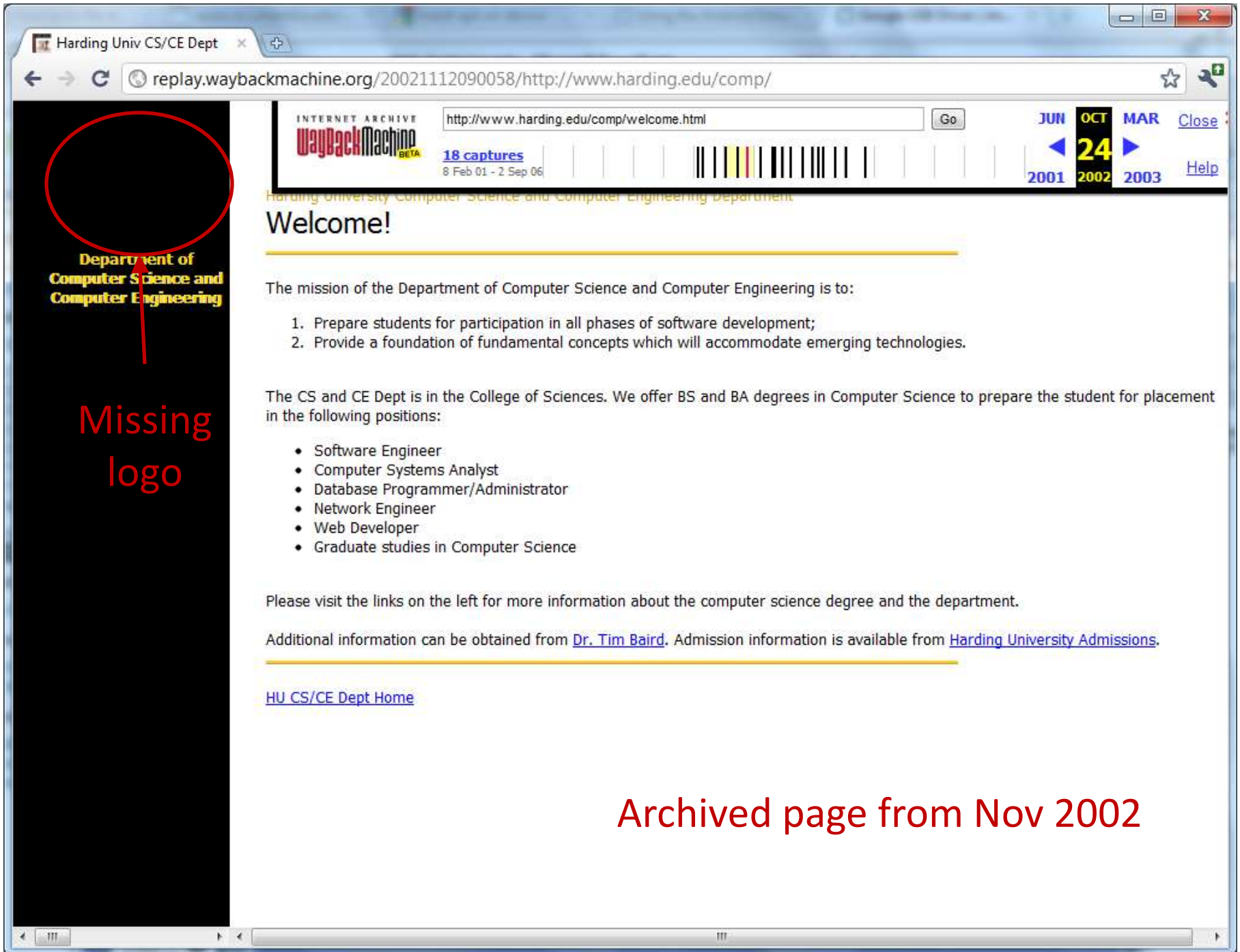
http://www.harding.edu/comp/

Go Wayback!

http://www.harding.edu/comp/ has been crawled 81 times going all the way back to November 12, 2002. A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ



JAN							FEB							MAR							APR																
				1	2						1	2	3	4	5	6					1	2	3	4	5	6									1	2	3
3	4	5	6	7	8	9	7	8	9	10	11	12	13	7	8	9	10	11	12	13	4	5	6	7	8	9	10										
10	11	12	13	14	15	16	14	15	16	17	18	19	20	14	15	16	17	18	19	20	11	12	13	14	15	16	17										
17	18	19	20	21	22	23	21	22	23	24	25	26	27	21	22	23	24	25	26	27	18	19	20	21	22	23	24										
24	25	26	27	28	29	30	28	28	29	30	31	25	26	27	28	29	30																				
31																																					
MAY							JUN							JUL							AUG																
						1							1	2	3	4	5								1	2	3	4	5	6	7						
2	3	4	5	6	7	8	6	7	8	9	10	11	12	4	5	6	7	8	9	10	8	9	10	11	12	13	14										
9	10	11	12	13	14	15	13	14	15	16	17	18	19	11	12	13	14	15	16	17	15	16	17	18	19	20	21										
16	17	18	19	20	21	22	20	21	22	23	24	25	26	18	19	20	21	22	23	24	22	23	24	25	26	27	28										
23	24	25	26	27	28	29	27	28	29	30	25	26	27	28	29	30	31	29	30	31																	
30	31																																				
SEP							OCT							NOV							DEC																



Department of
Computer Science and
Computer Engineering

Missing
logo

Welcome!

The mission of the Department of Computer Science and Computer Engineering is to:

1. Prepare students for participation in all phases of software development;
2. Provide a foundation of fundamental concepts which will accommodate emerging technologies.

The CS and CE Dept is in the College of Sciences. We offer BS and BA degrees in Computer Science to prepare the student for placement in the following positions:

- Software Engineer
- Computer Systems Analyst
- Database Programmer/Administrator
- Network Engineer
- Web Developer
- Graduate studies in Computer Science

Please visit the links on the left for more information about the computer science degree and the department.

Additional information can be obtained from [Dr. Tim Baird](#). Admission information is available from [Harding University Admissions](#).

[HU CS/CE Dept Home](#)

Archived page from Nov 2002



Web **Moving Images** Texts Audio Software Patron Info About IA Projects

Home Animation & Cartoons | Arts & Music | Community Video | Computers & Technology | Cultural & Academic Films | Ephemeral Films | Movies | News & Public Affairs | [Prelinger Archives](#) | Spirituality & Religion | Sports Videos | Videogame Videos | Vlogs | Youth Media

Search: Prelinger Archives Advanced Search **Anonymous User** (login or join us)

[Moving Image Archive](#) > [Prelinger Archives](#) > [Make Mine Freedom](#)

View movie



[View thumbnails](#)
Run time: 9:30

Stream [\(help ?\)](#)

[64Kb Real Media](#) (dialup)
[256Kb Real Media](#)
(broadband)

Play / Download [\(help ?\)](#)

- [Cinepack](#) (29.6 M)
- [Ogg Video](#) (38.4 M)
- [512Kb MPEG4](#) (38.6 M)
- [64Kb Real Media](#) (39.8 M)
- [256Kb Real Media](#) (94.7 M)
- [HiRes MPEG4](#) (103.8 M)
- [MPEG2](#) (251.1 M)

All Files: [HTTP](#)



Make Mine Freedom (1948)

Sutherland (John) Productions



[embed this](#)

Your browser supports the new <video> tag!
Would you like to [try the new <video> tag?](#)

This Cold War-era cartoon uses humor to tout the dangers of Communism and the benefits of capitalism.

This movie is part of the collection: [Prelinger Archives](#)

Producer: Sutherland (John) Productions

Sponsor: Harding College (Searcy, Arkansas)

Audio/Visual: Sd, C

Keywords: [Cold War](#); [Animation](#); [Advertising](#); [Capitalism](#)

Creative Commons license: [Public Domain](#)

Other Players

- National libraries & national archives, usually focusing on culturally significant web collections
 - US Library of Congress: [Minerva](#)
 - UK Web Archiving Consortium: [UK Web Archive](#)
 - National Library of Australia: [PANDORA](#)
 - Etc.
- Commercial organizations
 - Hanzo Archives: commercial web archiving tools
 - Nextpoint: archiving service for organizations
 - Iterasi: for corporate, legal, and govt
 - Etc.

Library of Congress Web Archives *Minerva*

BROWSE | SEARCH |
TECHNICAL INFORMATION

LC Web Archives

Web Archives Available:

- [Crisis in Darfur, Sudan, Web Archive, 2006](#)
- [Iraq War 2003 Web Archive](#)
- [Law Library Legal Blawqs Web Archive](#)
- [Library of Congress Manuscript Division Archive of Organizational Web Sites](#)
- [Papal Transition 2005 Web Archive](#)
- [September 11, 2001 Web Archive](#)
- [Single Sites Web Archive](#)
- [United States 107th Congress Web Archive](#)
- [United States 108th Congress Web Archive](#)
- [United States Election 2000 Web Archive](#)
- [United States Election 2002 Web Archive](#)
- [United States Election 2004 Web Archive](#)
- [United States Election 2006 Web Archive](#)
- [Visual Image Web Sites Archive](#)



The Library of Congress Web Archives (LCWA) is composed of collections of archived web sites selected by subject specialists to represent web-based information on a designated topic. It is part of a continuing effort by the Library to evaluate, select, collect, catalog, provide access to, and preserve digital materials for future generations of researchers. The early development project for Web archives was called MINERVA.

LC Web Archives

Other Players

- Free on-demand archiving
 - [WebCite](#): for saving citable web resources
 - www.backupurl.com
 - www.freezepage.com



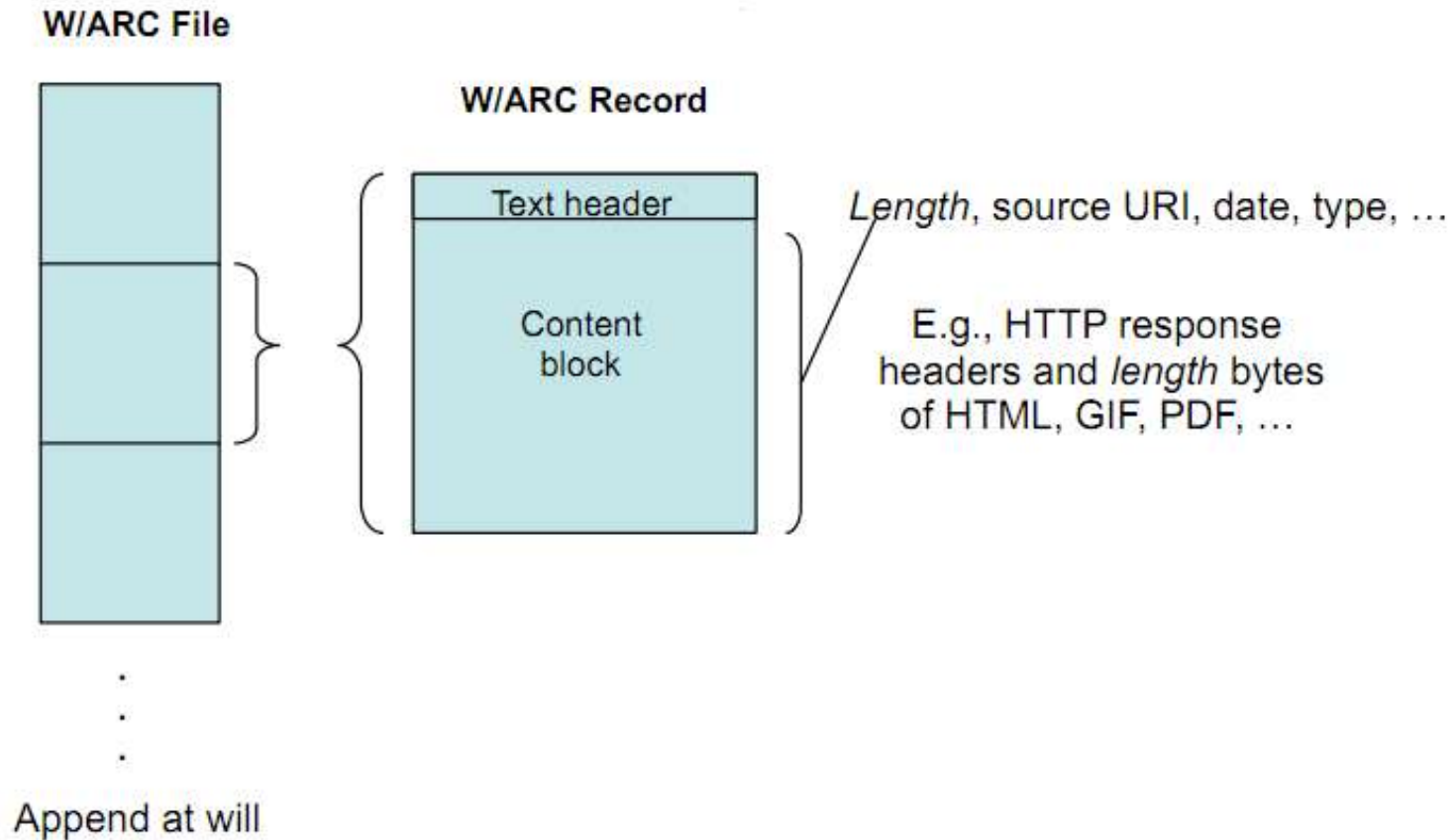
Special Collections

- Library of Congress has numerous collections
 - Twitter archive, 9/11, Iraq War, & much more
- Archive-it.org (ran by Internet Archive)
 - Homeless websites in LA, Virginia Tech, & much more
- Stanford WebBase
 - US Presidential election 2008, Virginia Tech shootings, Hurrikane Ike 2008
- Geocities archive
 - Number of archiving groups: [ReoCities](#), [OoCities](#), and Internet Archive
 - [652 GB torrent](#) also available
- ArchiveFacebook
 - Firefox add-on created to archive individual Facebook pages

Web Crawlers

- Wget and HTTrack
 - Simple tools for mirroring a website
 - All crawled URLs are converted into a path and file
 - `http://foo.org/test/` saved as `foo.org/test/index.html`
 - Not designed for large-scale crawling
- Heritrix
 - Built by Internet Archive and Nordic national libraries for larger web archiving tasks
 - Archived content stored in Web ARChive file format ([WARC](#))
 - Uses web interface
 - Can find links in JavaScript, Flash, etc.

WARC File Organization



Archiving the Deep Web

- How are Deep web websites archived when links aren't available to crawl?
- One strategy: Get website owner to release their database (legal deposit)
- [DeepArc](#) tool
 - Developed by National Library of France (BnF)
 - Transforms relational database content into XML for archiving purposes
- [Xinq](#) tool
 - Developed by National Library of Australia
 - Allows online browsing and searching of XML database

Let's explore some novel uses of
web archives



[Web](#) | [Moving Images](#) | [Texts](#) | [Audio](#) | [Software](#) |
[Patron Info](#) | [About IA](#)

Universal access
to human knowledge

[Home](#)

[Donate](#) | [Forums](#) | [FAQs](#) | [Contributions](#) | [Terms, Privacy, & Copyright](#) | [Contact](#) | [Jobs](#) | [Bios](#)

Search:

Forums 

Anonymous User [\(login or join us\)](#)

 [Advanced Search](#)

[View Post](#) [\[edit\]](#)

[Reply to this post](#) | [Go Back](#)

Poster: Ali Ross **Date:** October 25, 2004 08:18:32am
Forum: [web](#) **Subject:** Any way to download full sites?

Hi,

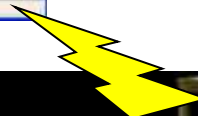
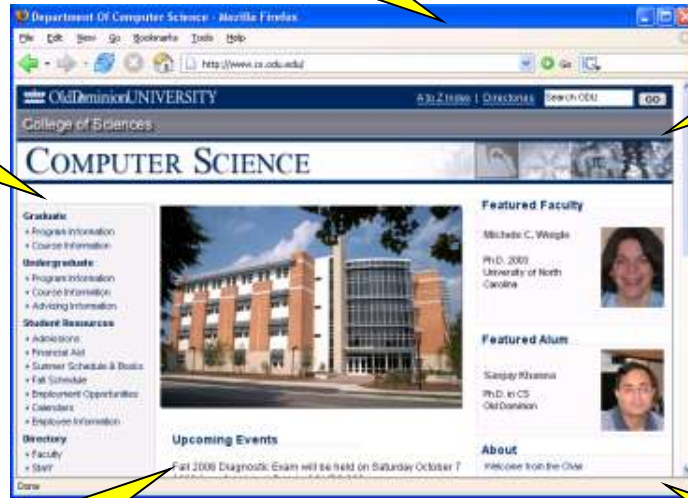
My old web hosting company lost my site in it's entirety (duh!) when a hard drive died on them. Needless to say that I was peeved, but I do notice that it is available to browse on the wayback machine.

I have used wget and such tools on linux to download full archives of sites in the past, but I can't seem to get anything other than index.html with wget. Does anyone have any ideas if I can download my full site?

Many thanks in advance,

Regards,

Alistair Ross.





Web Infrastructure

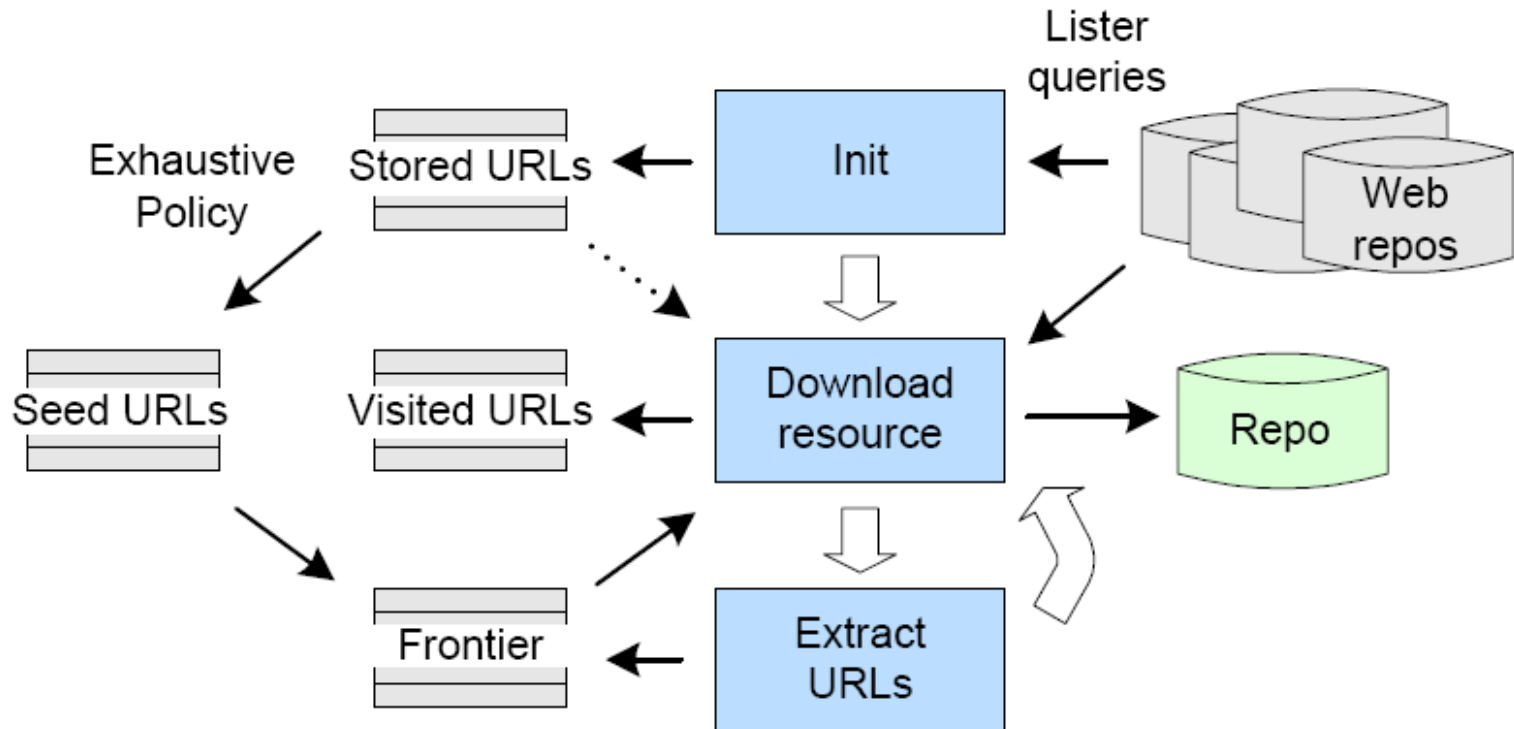


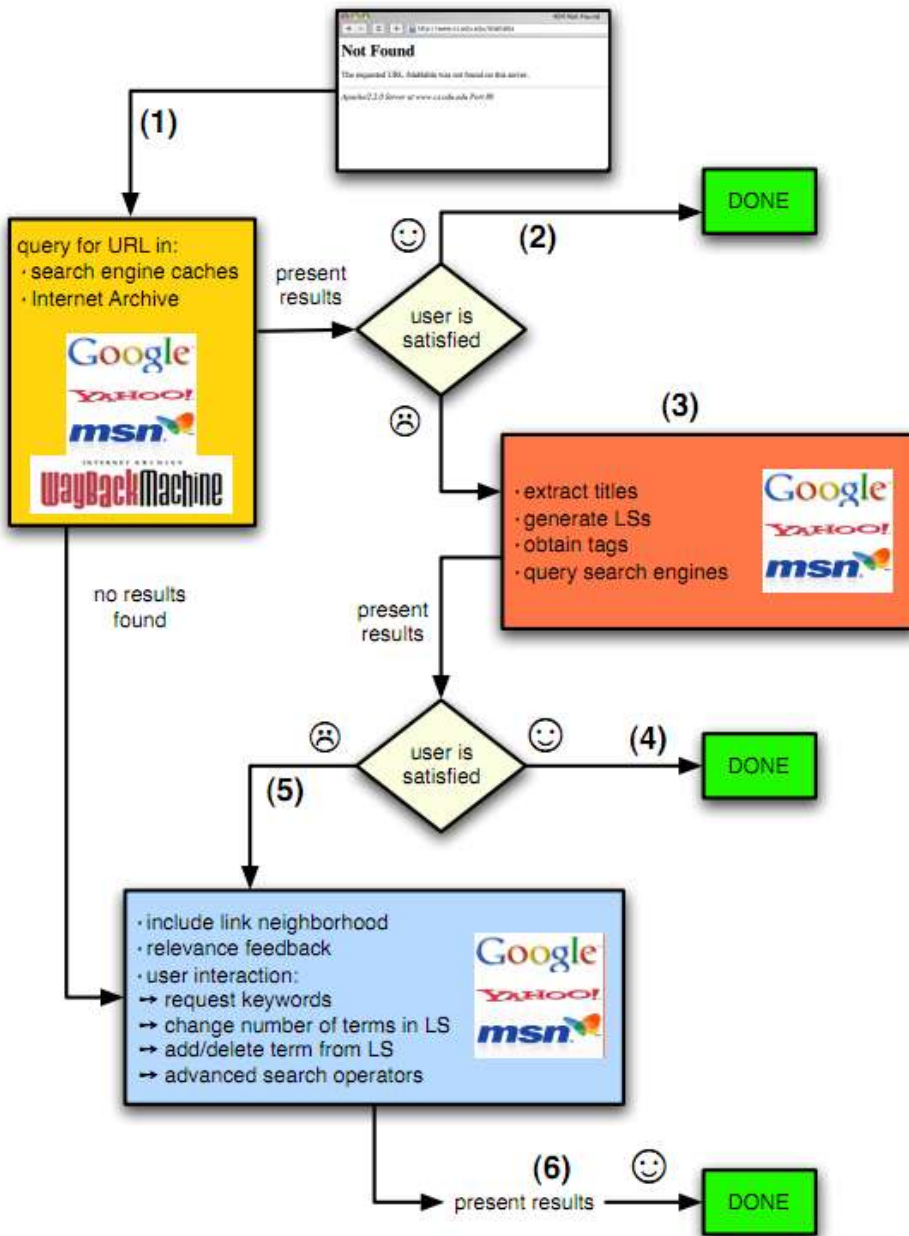
UK WEB ARCHIVING
CONSORTIUM
www.webarchive.org.uk



Web Repository Crawling

- Warrick developed in 2005 as a web repository crawler
- Used to recover thousands of websites from the WI
- Available at <http://warrick.cs.odu.edu/>





Using the WI to find missing web pages

Memento – Date/Time Negotiation

- Memento is a new protocol which uses HTTP content negotiation to retrieve older versions (Mementos) of web resources
- Agent request URI with Accept-Datetime set to desired date/time
- Server responds with a link to a TimeGate which knows the Mementos available for the URI
- Agent makes request to TimeGate and receives response with the URL to the Memento
- Learn more: <http://mementoweb.org/>

