Web Archiving Workshop

Mark Phillips

Texas Conference on Digital Libraries
June 4, 2008
Agenda

1:00 Welcome/Introductions
1:15 Introduction to Web Archiving
    History
    Concepts/Terms
    Examples
2:15 Collection Development Pt 1
3:00 Break
3:15 Collection Development Pt 2
4:00 Tools Overview
    Crawlers
    Display
    Search
4:45 Closing/Questions
Introductions

Name:
Institution:
Level: Beginner, Moderate, Advanced
Expectations: What did you want to leave today knowing that you don't already
Web Archiving History
(according to Mark)

1996 Internet Archive
1996 Pandora
1996 CyberCemetery
1997 Nordic Web Archive
2000 Library of Congress: Minerva
2001 International Web Archiving Workshop
2003 International Internet Preservation Consortium
2004 Heritrix first release
2005 NDIIPP Web-At-Risk project starts
2005 Archive-It released
2006 Web Curator Tool
2008 Web Archiving Service
Concepts/Terms

Crawler/Harvester/Spider
Seed
Robots Exclusion Protocol
Surts
Hops
Crawler Traps
Politeness
ARC/WARC
Crawler/Harvester/Spider

URL List

next URL in queue

URL Fetcher

Parser

add URL to queue

get URL

Web
Seed/entry-point URLs

A URL appearing in a seed list as one of the starting addresses a web crawler uses to capture content.

Seed for University of North Texas Web Harvest

http://www.unt.edu

Seeds for 2004 End of Term Harvest

2000+ URL's covering all branches of Federal government
Robots Exclusion Protocol

Sometimes referred to as robots.txt

Allows site owners to restrict content from being crawled by crawlers

It is common to obey robots.txt directives, though there are exceptions

http://whitehouse.gov (2100 lines long)

Image folders
Surt

Sort-friendly URI Reordering Transform

Normalizes URLs for sorting purposes

URL:
http://www.library.unt.edu
http://digital.library.unt.edu
http://www.unt.edu
https://texashistory.unt.edu
http://unt.edu

SURT:
http://edu,unt
http://edu,unt,library,digital
http://edu,unt,library,www
http://edu,unt,texashistory
http://edu,unt,www
Hops

A hop occurs when a crawler retrieves a URL based on a link found in another document.

Link Hop Count = Number of links follow from the seed to the current URI.

This is different than “Levels”
Levels

Websites don't really have “levels” or “depth”.

We think they do because all of us built a website at one time and we remember putting things in directories.

We sometimes see this
http://example.com/level1
http://example.com/level1/level2/level3/doc.html

But we also see this
http://example.com/page.php?ID=2
http://example.com/page.php?ID=55
hops

http://example.com

http://example.com/about

http://example.com/depts

http://example.com/depts/research

http://example.com/depts/research/new/
Crawler Traps

is a set of web pages that may intentionally or unintentionally cause a web crawler or search bot to make an infinite number of requests

Calendars that have next buttons forever
http://example.com/cal/year.php?year=2303

Content management systems
http://foo.com/bar/foo/bar/foo/bar/foo/bar/.....

Forums
Politeness

Politeness refers to attempts by the crawler software to limit load on a site.

adaptive politeness policy:

if it took $n$ seconds to download a document from a given server, the crawler waits for $2n$ seconds before downloading the next page.
arc/warc

arc: Historic file format used by the internet archive for storing harvested web content.

warc: New file format being developed by IIPC and others in attempts to create a standard for storing harvested web content.
Text format
Concatenated bitstreams into a series of larger files.
~100 MB per file containing many (thousands) of smaller files.
Contains metadata as well as content
The only way that web archives are able to store millions and billions of files on a standard file system.
arc vs warc

arc: older format, non standard, changed over time, more than 1.5 PB of legacy arc files

warc: new format, standard, developed by many people, allows for more metadata to be collected, allows for other uses for the warc file format. Should lead to adoption by tools like wget and Httrack

Heritrix has the option to write both formats.
Examples

- Internet Archive
- CyberCemetery
- Pandora
- Minerva
- Archive-it
Internet Archive

The Wayback Machine

[http:// Wayback Machine]

Advanced Search

Around the World in 2 Billion Pages

Thanks to a generous grant from the Mellon Foundation, Internet Archive completed a 2 billion page web crawl in 2007. This is the largest web crawl attempted by Internet Archive. The project was designed to take a global snapshot of the Web.

Please browse through the resulting collection.

Special thanks to the memory institutions who contributed URLs to the crawl. The crawl began with 18,000 websites from over 60 countries.

Web Archiving Services

[Archive-It]

Archive-It allows institutions to build and preserve their own web archive of born digital content, through a user friendly web application, without requiring any technical expertise or hosting facilities. Subscribers can harvest, catalog, and archive their collections, and then search and browse the collections when complete. Collections are hosted at the Internet Archive data center, and accessible to the public with full text search.

With the rapid growth of the Internet and the World Wide Web, millions of people have grown accustomed to using these tools as resources to acquire information, and the availability of electronic information is taken for granted. However, it is a fallacy that if something is on the web, it will be there forever. The average lifespan of a web page is 44–75 days. There's an urgent need for people to understand that that web is who we are. It's our culture and our social fabric, and we don't want to lose any of it. What is here today might be gone tomorrow.

Archive-It is designed to fit the needs of many types of organizations and individuals. The over 50 partners include state archives, university libraries, federal institutions, state libraries, non-government non-profits, museums, historians, and independent researchers.
The University of North Texas Libraries and the U.S. Government Printing Office, as part of the Federal Depository Library Program, created a partnership to provide permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions. This collection was named the “CyberCemetery” by early users of the site.

The CyberCemetery is an archive of government websites that have ceased operation (usually websites of defunct government agencies and commissions that have issued a final report). This collection features a variety of topics indicative of the broad nature of government information. In particular, this collection features websites that cover topics supporting the university’s curriculum and particular program strengths.

This collection supports the mission of the UNT Libraries by acquiring, presenting, and providing access to recorded knowledge in the form of government websites.
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.
Welcome to Archive-IT

- Start with a tour
- Attend a live online demo: June 10 at 11am PDT or June 24 at 11am PDT
- Watch a pre-recorded online demo now
- View all Archive-IT collections

Archive-IT's subscription service, Archive-IT, allows institutions to build and preserve collections of born digital content, through a user-friendly web application, without requiring any technical expertise or hosting facilities. Subscribers can harvest, catalog, manage, and archive their collections, and then search and browse the collections whenever they choose. Collections are hosted at the Internet Archive data center, and accessible to the public with full-text search. Our partner support provides a user manual, training, and guidance throughout your collection building process, with a Partner Support Specialist available to assist you.

Featured Collections

- View Archive-IT K-12 Web Archiving Pilot Collections

Rice University Web Sites

http://www.archive-it.org/collections/7703

Rice University web pages reflecting the daily business and mission of Rice University, in Houston, Texas.

International Institute for Social History: South East Asia

http://www.archive-it.org/collections/2117

Collection of websites of organizations from Burma/Myanmar, Cambodia, Malaysia, Thailand, concerned with labor issues, human rights, emancipation, environmental questions and other issues.

Utah State Library and Utah State Archives and Records Service: Natural Resources

http://www.archive-it.org/collections/528


Archive-IT is designed to fit the needs of many types of organizations and individuals. The over 50 partners include: state archives, university libraries, federal institutions, state libraries, non-government non-profits, museums, historians, and independent researchers.

We have some exciting developments scheduled for Archive-IT in 2008, including:

- A K-12 web archiving pilot project with schools around the country
- Spanish and Chinese User interface to encourage web archiving with global partners
- Crisis, Tragedy, and Recovery web crawl service for spontaneous collections

Learn More