The Great WARC Adventure

Nick Ruest (York University)
Ian Milligan (University of Waterloo)
Today’s Talk

* A brief historical overview of web archiving
* How to capture and preserve
* Discoverability and usability (with open source software)
* Interplay of the archivist and the historian
* Piecing the story together
* Internet Archive isn’t the only way
Historical overview of web archiving
Going back to 1995/1996

We do not know today what Mozart sounded like on the keyboard ... What will future generations know of our history? ... But digital technology seemed to come to the rescue, allowing indefinite storage without loss. Now we find that digital information too, has its dark side. (Michael Lesk, 1995)
Internet Archive Announces the "Wayback Machine"

Subject: Internet Archive Announces the "Wayback Machine"
From: david@nach.org

NINCH-ANNOUNCE <david@nach.org>
Subjects: Internet Archive Announces the "Wayback Machine"

---

Internet Archive Launches Wayback Machine

A fascinating and useful tool has just been unveiled by the Internet Archive, enabling the public to look back over any of the 10 billion web pages archived during its sweeps of the Internet since 1996. Type in a URL and see what the Wayback machine produces.

---

David Green

----------

Date: Thu, 25 Oct 2001 13:16:08 -0400 (EDT)
From: Ann Dershowitz, Ann.dershowitz@yale.edu
To: cni-announce@lists.nach.org
Subject: Archaeology, way back to - well for it - "1996"

Forwarded message

Jack Lynch wrote:

Date: Wed, 24 Oct 2001 19:38:19 -0400 (EDT)
From: Jack Lynch, jlynch@andromeda.rutgers.edu
Subject: Archaeology, way back to - well for it - "1996" (I'm told)

INTERNET ARCHIVE LAUNCHES WAYBACK MACHINE

Free Service Enables Users to Access Archived Versions of Web Sites Dating from 1996

SAN FRANCISCO (October 24, 2001) < The Internet Archive, a comprehensive library of Internet sites and cultural artifacts in digital form,
But that’s not all...
and... [http://freedaleaskey.plggta.org/]
How to capture and preserve
Progressive Librarians Guild
Toronto Area Chapter

PLG GTA Stands With Dale Askey

Posted on February 6, 2013 by plgta

As Toronto-area library workers concerned with issues of freedom of expression, censorship and freedom of information, we wish to issue a statement of support for Dale Askey, AUL at McMaster University, against the unprecedented libel lawsuit brought against him by the Edwin Mellen Press.

As a professional librarian engaged in collection development, Mr. Askey is qualified to make a judgment on the quality of published material. Furthermore, in an open and democratic society, he is free to share these opinions as he sees fit. The lawsuit brought against him is nothing more than a retaliatory measure and an outright assault on academic freedom—a principle that is highly regarded by both the doctoral community and libraries—the very communities that Edwin Mellen Press claims as their prime audience.

We call upon the library and academic communities to stand up for Dale’s academic freedom and basic freedom of speech. The Canadian Association of University Teachers has said the following about librarians and academic freedom:

Librarians have a duty to promote and preserve intellectual freedom in society. They have a responsibility to protect academic freedom and are entitled to the full protection of their own academic freedom in accordance with CAUT policies. This freedom includes, but is not limited to, the right and duty to exercise their academic professional judgment in the selection of library materials, and to ensure that library materials are freely accessible to all, no matter how controversial those materials may be.

Both the suit against Askey and past interactions with the scholarly community suggest Mellen Press is not interested in meaningful dialogue about scholarship and scholarly publishing. Currently, libraries are amongst the biggest customers of Edwin Mellen Press. Going forward, we hope to see a change in the relationship between Edwin Mellen Press and the community it claims to serve. In our opinion, an attack upon the academic freedom of one librarian is an attack upon us all.

This work, unless otherwise expressly stated, is licensed under a Creative Commons Attribution-ShareAlike 2.5 Canada License.
Terminal

Opening WARC file 'boycottmellenpress.warc'.

Resolving boycottmellenpress.com (boycottmellenpress.com)... 50.23.239.98
Connecting to boycottmellenpress.com (boycottmellenpress.com)[50.23.239.98]:80... connected.
HTTP request sent, waiting response... 200 OK
Length: 3004 (2.9K) [text/html]
Saving to: 'index.html'

2014-06-01 10:12:54 (307 MB/s) - 'index.html' saved [3004/3004]

[nuestub-SC-51840B:aca]$ cat boycottmellenpress.warc

WARC/1.0
WARC-Type: warcinfo
Content-Type: application/warc-fields
WARC-Date: 2014-06-03T20:12:52Z
WARC-Record-ID: <urn:uuid:0b0b7a9b-f3ea73-431a-9fa4-6fadebca400b>
WARC-Fileset: boycottmellenpress.warc
WARC-Block-Digest: sha1:4986c3393f2618b6c197c13c10f6c3e0
Content-Lengths: 208

software: Wget/1.14 (linux-gnu)
format: WARC File Format 1.0
conformance: http://bitstream.lboro.ac.uk/WARC/WARC_ISO_28500_version1_latestdraft.pdf
robots: classic

WARC/1.0
WARC-Type: request
WARC-Target-URL: http://boycottmellenpress.com/
Content-Type: application/http/msgtype=request
WARC-Date: 2014-06-01T12:23:42Z
WARC-Record-ID: <urn:uuid:3h55468b-28c1-479d-81e4-0239c61f70>
WARC-IP-Address: 50.23.239.98
WARC-WarcInfo-ID: <urn:uuid:280b5783f-6e73-431a-9fa4-c6e8bca400b>
WARC-Block-Digest: sha1:F7GN2P4XZ53SEAEUVYOVX53BDCM07U
Content-Lengths: 128

GET / HTTP/1.1
User-Agent: Wget/1.14 (linux-gnu)
Accept: */
Host: boycottmellenpress.com
Connection: Keep-Alive

WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:b70b7e89-5699-4d0b-bd00-b20816258769>
WARC-WarcInfo-ID: <urn:uuid:8989783f-6e73-431a-9fa4-36e8bca400b>
WARC?
PROVENANCE!
<h2>Are we still boycotting</h2>
<h2>Edwin Mellen Press</h2>
<h2>YES!!</h2>
<h2>YES!!</h2>
<h2>Why?</h2>
TOOLS AND SOFTWARE

In the perspective of setting up a Web archiving chain, the following tools are recommended and used by members of the IPC:

**Acquisition**

ArchiveIt Facebook, a Mozilla Firefox add-on for individuals to archive their Facebook accounts
Developed by: Mat Kelly, Carlton Nordhem, Harry Salmon, Michael Nelson, and Frank McCown
Current version: 1.4

Heritrix, an open source, extensible, web-scale, archival quality web crawler
Developed by: Internet Archive with the Nordic national libraries
Current versions: Heritrix 3.1.1 (2012-05-02); Heritrix 1.14.4 (2010-05-10) and Heritrix 0.2.2 (2009-11-08)
More information: https://webarchive.jira.com/wiki/display/Heritrix/heritrix
Download (3.1.x): http://builds.archive.org/3080/maven2/org/archive/heritrix/heritrix/
Download (2.1.x): http://sourceforge.net/projects/archive-crawler/

HTTrack, an open source website copying utility
Developed by: Xavier Roche and other contributors
Current version: 3.46-1 (2012-06-23)
More information: http://www.httrack.com/

SiteStory, a transactional archive that selectively captures and stores transactions that take place between a web client (browser) and a web server
Developed by: Los Alamos National Laboratory
Current version: 1.0
Download: http://mementoweb.github.com/SiteStory/

WARCreate, a Google Chrome extension for archiving an individual webpage or website to a WARC file
Developed by: Mat Kelly
Current version: unreleased
More information: http://matkelly.com/warcreate/

Warrick, an open source downloadable tool or web service for reconstructing websites from web archives, using Memento
Developed by: Frank McCown
Current version: 2.2.1 (2012-04)
More information: http://warrick.cs.ualberta.ca/
Download: http://code.google.com/p/warrick/downloads/list

Wget, an open source file retrieval utility
Current version: 1.14 (2012-08-05)
Light weight

wget --mirror --page-requisites --warc-file=THIS_AWESOME_SITE http://thisawesomesite.ca
Industrial strength

Heritrix
Discoverable and usable using open source software
Can we do this with Open Source software?
Use the tools you know

Drupal
Fedora Commons
Islandora
Heritrix
wkhtmltopdf
wkhtmltoimage
Can we talk OAIS?
SIP - Submission Information Package

Heritrix
Wget
wkhtmltopdf
wkhtmltoimage

...and Bash!
AIP - Archival Information Package

Drupal
Islandora
Fedora Commons
Web ARChive SP
Checksum
Checksum checker
Premis
FITS
DIP - Dissemination Information Package

Drupal
Islandora
Fedora Commons
Web ARChive SP
Interplay of the archivist and historian
An Ideal Case Study?
A Historian in the Archive

* WARC-Tools (https://code.google.com/p/warc-tools/)
* By Date (awesome!)
A Historian in the Archive

* Don’t read the comments

* Disqus
A Historian in the Archive
Learning from Word Frequency
A Historian in the Archive (distant reading)
A Historian in the Archive (distant reading)
A Historian in the Archive (distant reading)
Keywords = Gotta Know What You’re Looking For
What was the buzz during the Ontario leader's debate?

Toronto Star - 5 hours ago
Although she was on the defensive from the outset of the debate, an informal poll of Star readers showed 40 per cent thought Liberal Leader ...

Ontario leaders debate: Gas-plant scandal, jobs take centre stage
CBC.ca - 17 hours ago

Corruption, bad math, gas plants the focus of Ontario leaders' debate
Globalnews.ca - 10 hours ago

From zingers to math woes: Four highlights from the Ontario debate
The Globe and Mail - 4 hours ago

Gas plant scandal, Million Jobs Plan dominate Ontario leaders' debate
CANOE - 13 hours ago

Explore in depth (638 more articles)
A Historian in the Archive (distant reading) - search ‘librarians’
A Historian in the Archive (distant reading) - search ‘dale askey’
Problem: still need to know what you’re looking for!
A Historian in the Archive (distant reading)
A Historian in the Archive (distant reading)
Helps to piece the story together from massive web archives
Internet Archive isn’t the only way!
...but they created the Web Archiving Lifecycle
Thanks!

Nick Ruest: ruestn@yorku.ca / @ruebot
Ian Milligan: i2milligan@uwaterloo.ca / @ianmilligan1