

# 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)



ARCHIVE



INTERNET ARCHIVE




INTER

# The Internet Archive

old.cni.org/hforums/ninch-announce/2001/0131.html

Keep W Lib HistD Z / GM W AH.ca MMA RSS UTF8 YShout! Globe Open Access Butt... ArchiveHub - Res...



**CNI has a new website: [www.cni.org](http://www.cni.org)**

This page, from the old web site, describes past work. We are in the process of moving older material to the new site.

[Home](#)

[About CNI](#)

[Program Plan](#)

[Browse By Topic](#)

[Meetings & Events](#)

[Resources](#)

[Contact Us](#)

## Internet Archive Announces the "Wayback Machine"

**Subject:** Internet Archive Announces the "Wayback Machine"  
**NINCH-ANNOUNCE** ([david@ninch.org](mailto:david@ninch.org))  
**Date:** Thu, 25 Oct 2001 14:54:27 -0400

• Messages sorted by: [[date](#)] [[thread](#)] [[subject](#)] [[author](#)]

Message-Id: <p05100304b7fe0ec64b56@192.100.21.22]>  
Date: Thu, 25 Oct 2001 14:54:27 -0400  
To: ninch-announce@ninch.org  
From: NINCH-ANNOUNCE <david@ninch.org>  
Subject: Internet Archive Announces the "Wayback Machine"

NINCH ANNOUNCEMENT  
News on Networked Cultural Heritage Resources  
from across the Community  
October 25, 2001

INTERNET ARCHIVE LAUNCHES WAYBACK MACHINE  
<http://web.archive.org/>

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 Okerson <ann.okerson@yale.edu>  
>To: liblicense-l@lists.yale.edu  
>Subject: Archeology, way back to - wait 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: Archeology, way back to - wait for it! - \*1996! (fwd)

INTERNET ARCHIVE LAUNCHES WAYBACK MACHINE

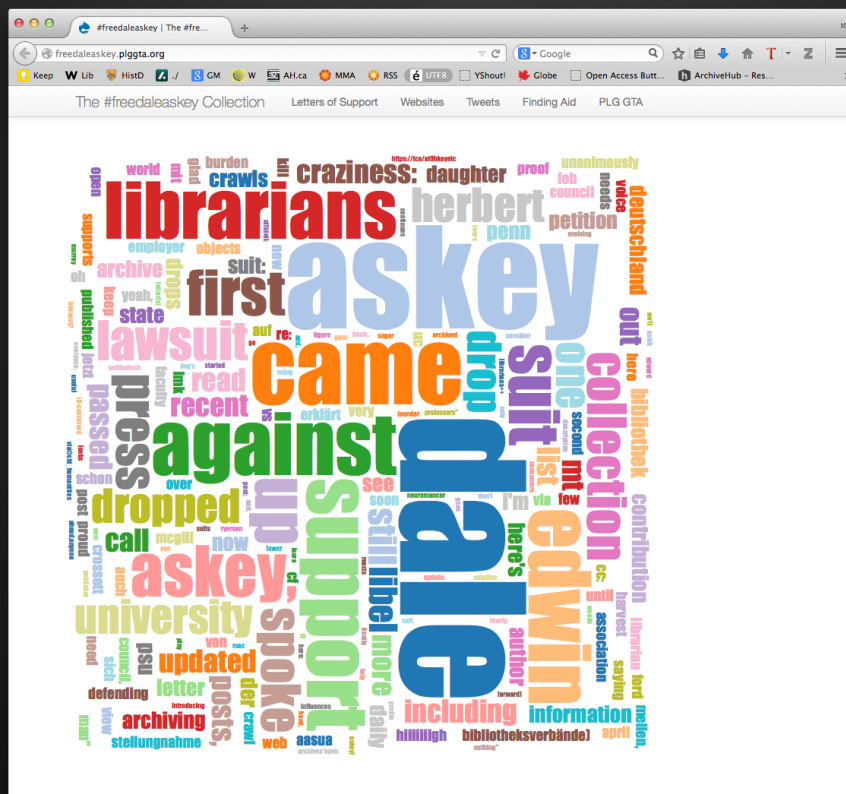
Free Service Enables Users to Access Archived Versions of Web Sites Dating from 1996

SAN FRANCISCO (October 24, 2001) < The Internet Archive, a comprehensive library of Internet sites and other cultural artifacts in digital form,





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 8, 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](#).

### Pages

- [About](#)
- [Contact](#)
- [How to Join](#)

### Links

- [#freedaleaskey Collection](#)
- [PLG Edmonton](#)
- [PLG London](#)
- [Progressive Librarians Guild](#)
- [Zotero group](#)

### Categories

- [Communication](#)
- [Discussion](#)
- [Statements](#)
- [Uncategorized](#)

### Archives

- [February 2014](#)
- [January 2014](#)
- [December 2013](#)
- [November 2013](#)
- [October 2013](#)
- [August 2013](#)
- [May 2013](#)
- [March 2013](#)
- [February 2013](#)
- [January 2013](#)
- [November 2012](#)
- [October 2012](#)
- [August 2012](#)
- [July 2012](#)

```
Terminal
[nruest@LB-SC-S1B4DB:aca]$ wget --warc-file=boycottmellenpress --no-warc-compression http://boycottmellenpress.com
Opening WARC file 'boycottmellenpress.warc'.

--2014-06-03 16:12:53-- http://boycottmellenpress.com/
Resolving boycottmellenpress.com (boycottmellenpress.com)... 50.23.239.98
Connecting to boycottmellenpress.com (boycottmellenpress.com)[50.23.239.98]:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3004 (2.9K) [text/html]
Saving to: 'index.html'

  OK ..                               100% 307M=0s

2014-06-03 16:12:54 (307 MB/s) - 'index.html' saved [3004/3004]

[nruest@LB-SC-S1B4DB:aca]$ cat boycottmellenpress.warc
WARC/1.0
WARC-Type: warcinfo
Content-Type: application/warc-fields
WARC-Date: 2014-06-03T20:12:53Z
WARC-Record-ID: <urn:uuid:809b783f-ea73-431a-9fa4-c6e0bcc40bbc>
WARC-Filename: boycottmellenpress.warc
WARC-Block-Digest: sha1:RMV6QLTEU6ECCRKM6SB6GLJUPBPQVFBM
Content-Length: 268

software: Wget/1.14 (linux-gnu)
format: WARC File Format 1.0
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf
robots: classic
wget-arguments: "--warc-file=boycottmellenpress" "--no-warc-compression" "http://boycottmellenpress.com"

WARC/1.0
WARC-Type: request
WARC-Target-URI: http://boycottmellenpress.com/
Content-Type: application/http;msgtype=request
WARC-Date: 2014-06-03T20:12:54Z
WARC-Record-ID: <urn:uuid:256cae8b-28c1-479d-b1e4-e02a9c61f70f>
WARC-IP-Address: 50.23.239.98
WARC-Warcinfo-ID: <urn:uuid:809b783f-ea73-431a-9fa4-c6e0bcc40bbc>
WARC-Block-Digest: sha1:F7GN2MHQRZ35EAZEUVYVIVIXSJBCHQD7U
Content-Length: 120

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:b70b7e89-5609-4d0b-bd00-b20816258769>
WARC-Warcinfo-ID: <urn:uuid:809b783f-ea73-431a-9fa4-c6e0bcc40bbc>
```



**WARC?**

**PROVENANCE!**

```

WARC/1.0^M
WARC-Type: warcinfo^M
Content-Type: application/warc-fields^M
WARC-Date: 2014-05-02T03:39:01Z^M
WARC-Record-ID: <urn:uuid:1b530dd3-df05-4ffa-8843-fec84fd727f3>^M
WARC-Filename: boycottmellenpress-2014_05_01.warc^M
WARC-Block-Digest: sha1:TP3MEJQ7DVS36J0ULUEE3DINUMKC36II^M
Content-Length: 326^M
^M
software: Wget/1.14 (linux-gnu)^M
format: WARC File Format 1.0^M
conformsTo: http://bibnum.bnf.fr/WARC/WARC_ISO_28500_version1_latestdraft.pdf^M
robots: classic^M
wget-arguments: "--mirror" "--page-requisites" "--warc-file=boycottmellenpress-2014_05_01" "--no-warc-compression" "--wait=1" "http://www.boycottmellenpress.com/" ^M
^M
^M
WARC/1.0^M
WARC-Type: request^M
WARC-Target-URI: http://www.boycottmellenpress.com/^M
Content-Type: application/http;msgtype=request^M
WARC-Date: 2014-05-02T03:39:02Z^M
WARC-Record-ID: <urn:uuid:a0e576b0-169b-499f-be3a-e963a09ba1f4>^M
WARC-IP-Address: 50.23.239.98^M
WARC-WarcInfo-ID: <urn:uuid:1b530dd3-df05-4ffa-8843-fec84fd727f3>^M
WARC-Block-Digest: sha1:QR775I05R263VLKHSCLKM5QQPHW4RE2G^M
Content-Length: 124^M
^M
GET / HTTP/1.1^M
User-Agent: Wget/1.14 (linux-gnu)^M
Accept: */*^M
Host: www.boycottmellenpress.com^M
Connection: Keep-Alive^M
^M
^M
WARC/1.0^M
WARC-Type: response^M
WARC-Record-ID: <urn:uuid:47fc1fd9-4234-4585-b8fa-35c7e93c0901>^M
WARC-WarcInfo-ID: <urn:uuid:1b530dd3-df05-4ffa-8843-fec84fd727f3>^M
WARC-Concurrent-To: <urn:uuid:a0e576b0-169b-499f-be3a-e963a09ba1f4>^M
WARC-Target-URI: http://www.boycottmellenpress.com/^M
WARC-Date: 2014-05-02T03:39:02Z^M
WARC-IP-Address: 50.23.239.98^M
WARC-Block-Digest: sha1:KERCJNX2NCHVYN5CCY7D5VBHOXW4QE2Y^M
WARC-Payload-Digest: sha1:PP6HSHOIZJL47D2UMVB4ZCK3QC7CUVJM^M
Content-Type: application/http;msgtype=response^M
Content-Length: 3352^M
^M
HTTP/1.1 200 OK^M
Date: Fri, 02 May 2014 03:39:02 GMT^M
Server: Apache^M
Last-Modified: Wed, 12 Jun 2013 18:41:15 GMT^M
ETag: "39947c5-bbc-4def95bc988c0"

```

```
Content-Length: 3004^M
Cache-Control: max-age=86400^M
Expires: Sat, 03 May 2014 03:39:02 GMT^M
Keep-Alive: timeout=2, max=100^M
Connection: Keep-Alive^M
Content-Type: text/html^M
^M
<!DOCTYPE html>
<html class=no-js manifest=a.appcache>
<head>
<title>Boycott Edwin Mellen Press</title>
<meta name=description content="Boycott Edwin Mellen Press">
<meta name=viewport content="width=device-width">
<link href='http://fonts.googleapis.com/css?family=Open+Sans' rel='stylesheet' type='text/css'>
<link href='http://fonts.googleapis.com/css?family=Gentium+Book+Basic' rel='stylesheet' type='text/css'>
<style type="text/css">
  body {
    -webkit-font-smoothing: antialiased;
    padding-top: 2.5%;
    padding-right: 15%;
    padding-left: 15%;
    padding-bottom: 10%;
    text-align: center;
    background: #EF4723;
    color: #222;
    min-height: 100%;
  }
  h1 {
    text-transform: uppercase;
    margin: 0 0 32px 0;
    text-align: center;
    font: 800 128px/128px "open sans";
  }
  h2 {
    text-transform: uppercase;
    font: 800 48px/48px "open sans";
  }
  p {
    text-align: center;
    font: 400 18px/18px "gentium book basic";
  }
  a, a:visited {
    color: #eee;
  }
</style>
</head>
<body>
<h2>Are we still boycotting</h2>
<h2>Edwin Mellen Press?</h2>
<h1>YES</h1>
<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 IIPC:

### Acquisition

**ArchiveFacebook**, a [Mozilla Firefox](#) add-on for individuals to archive their Facebook accounts  
Developed by: Mat Kelly, Carlton Northern, Hany SalahEldeen, Michael Nelson, and Frank McCown  
Current version: 1.4  
More information: <https://addons.mozilla.org/en-US/firefox/addon/archivefacebook/>

**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 2.0.2 (2008-11-08)  
More information: <https://webarchive.iira.com/wiki/display/Heritrix/Heritrix>  
Download (3.X): <http://builds.archive.org:8080/maven2/org/archive/heritrix/heritrix>  
Download (2.X, 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  
More information: <http://www.dlib.org/dlib/september12/09inbrief.html>  
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.odu.edu/>  
Download: <http://code.google.com/p/warrick/downloads/list>

**Wget**, an open source file retrieval utility  
Current version: 1.14 (2012-08-05)  
More information: <http://www.gnu.org/software/wget/>, [http://www.archiveteam.org/index.php?title=Wget\\_with\\_WARC\\_output](http://www.archiveteam.org/index.php?title=Wget_with_WARC_output)  
Download: <ftp://ftp.gnu.org/gnu/wget/>

### NEWS

- NATIONAL LIBRARY OF SCOTLAND RE-JOINS IIPC
- REGISTRATION FOR PARIS GENERAL ASSEMBLY NOW CLOSED
- REGISTRATION FOR PARIS GENERAL ASSEMBLY NOW OPEN

RT [@C\\_Fryer](#): Thanks to [@anjacks0n](#) for having a go at What is Preservation Planning? | Q&A Answered | <http://t.co/vH4RjYOdly> [#practicaldigip...](#)

3 hours 8 min ago



Follow Us

# Light weight

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

# Industrial strength

Heritrix

# Discoverable and usable

using open source software



[HOME](#) [EXPLORE](#) [LEARN MORE](#) [CONTACT US](#)

The leading web archiving service  
for collecting and accessing  
cultural heritage on the web  
*Built at the Internet Archive*



Login

Welcome to Archive-It!  
Attend a live informational webinar and demo  
to learn more about the service

Contact Us to sign up for an upcoming session:  
Jun 17 2014, 11:30 AM PDT  
Jul 08 2014, 11:30 AM PDT

Explore Collections

[Show All Collections](#)



### Earthquake in Haiti

By Internet Archive Global Events

This collection is currently documenting the events of the January 2010 earthquake in Haiti and the aftermath, including the rescue efforts from around the world and...



### Wikileaks Document Release Collection

By Internet Archive Global Events

A collection of websites, news coverage, and commentary surrounding the Wikileaks releases starting in 2010. Includes documents released from the Afghan war diaries,...



### Jasmine Revolution - Tunisia 2011

By Internet Archive Global Events

A collection of websites, news coverage, and commentary surrounding the 2011 Jasmine Revolution in Tunisia. Our partners at Library of Congress and Bibliothèque Nationale de...

Explore Collecting Organizations

[Show All Organizations](#)



### Winthrop University

Founded in 1886, Winthrop University is a public, coeducational, comprehensive teaching university that teaches students to live, learn, and lead for a lifetime. The university...



### University of Wisconsin

This collection currently includes two distinct sub-collections: The UW-Madison Collection and The Stem Cell Research Archives Project. The UW-Madison Collection...



### Columbia College Chicago

Columbia College Chicago is an international leader and recognized pioneer in arts and media education. With over 120 years of deep experience teaching creative students to...

\$\$\$

**Can we do this with Open  
Source software?**

# Use the tools you know

Drupal

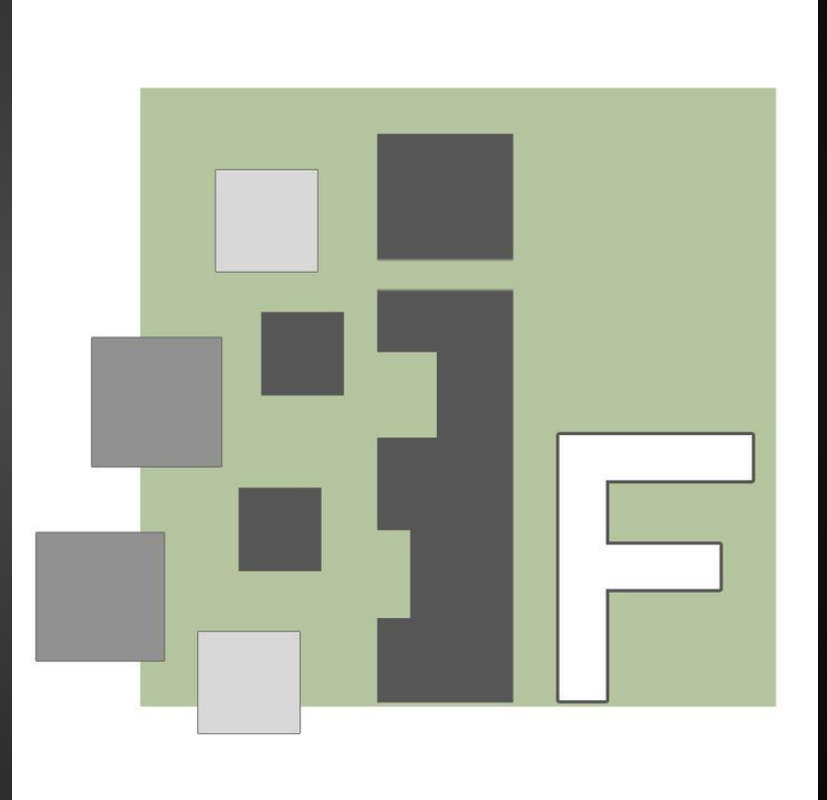
Fedora Commons

Islandora

Heritrix

wkhtmltopdf

wkhtmltoimage





**Can we talk OAIS?**

# SIP - Submission Information Package

Heritrix  
Wget  
wkhtmltopdf  
wkhtmltoimage

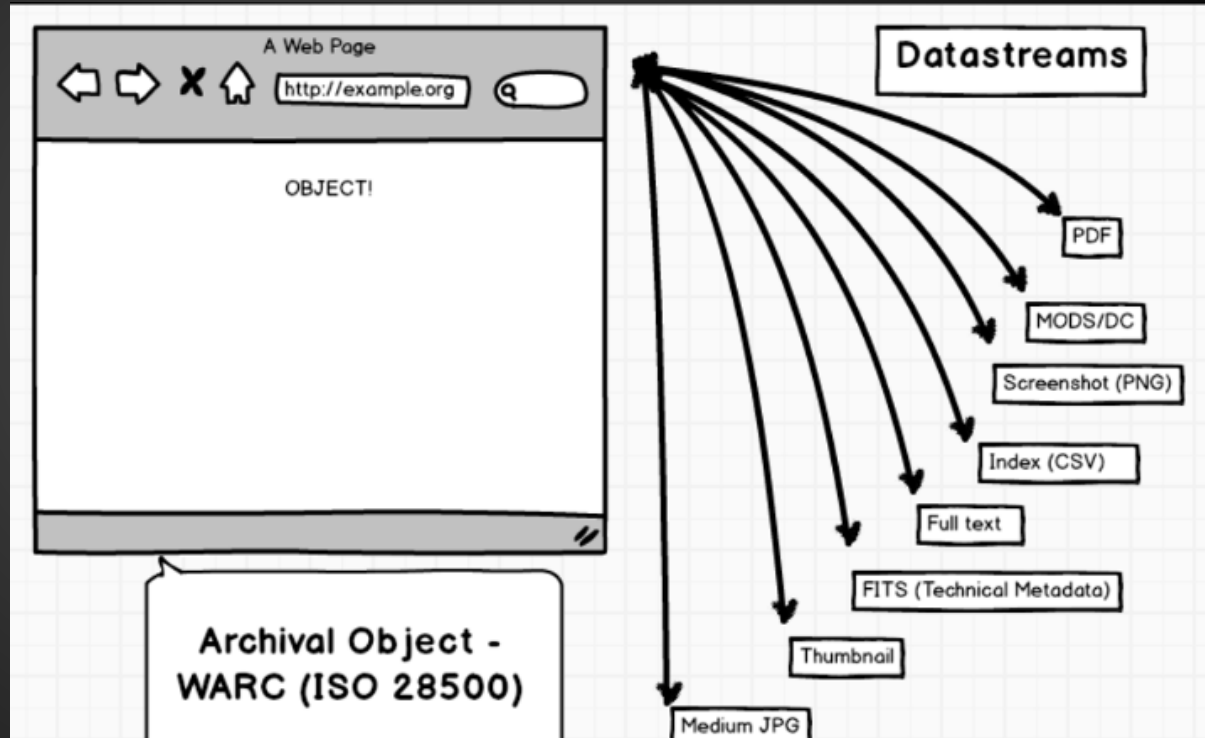
...and Bash!

```
1 #!/bin/bash
2 # Bash script that captures selected web sites using the WARC ISO format.
3
4 usage ()
5 {
6     cat <<EOF
7     Usage: $0 options
8
9     Program to capture websites.
10
11     OPTIONS:
12     -h Show this message
13     -t Title/name of the website being captured
14     -u URL of the website being captured
15 EOF
16 }
17
18 DATE=$(date +%Y_%m_%d)
19
20 while getopts ":ht:u:" OPTION
21 do
22     case $OPTION in
23         h)
24             usage
25             exit 1
26             ;;
27         t)
28             TITLE=$OPTARG
29             ;;
30         u)
31             URL=$OPTARG
32             ;;
33     esac
34 done
35
36 printf "Creating a pdf of: $URL \n"
37 xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltopdf $URL $DATE.pdf
38 printf "Created: $TITLE-$DATE.pdf \n"
39 printf "Taking a screenshot of: $URL \n"
40 xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltoimage $URL $DATE.png
41 pngcrush temp.png $TITLE-$DATE.png
42 rm temp.png
43 printf "Created: $TITLE-$DATE.png \n"
44 printf "Crawling: $URL \n"
45 /usr/local/bin/wget --mirror --page-requisites --convert-links --no-parent --random-wait --warc-file=$DATE.warc
46 printf "Created: $TITLE-$DATE.warc"
```

```
1 #!/bin/bash
2 # Bash script that archives selected Free Dale Askey web sites using the WARC ISO format.
3 HOME=/mnt/DIY/web-archiving/arkivdaleascii/sites
4 DATE=$(date +%Y_%m_%d)
5 SITES=/home/nruerst/git/arkivdaleascii/arkivdaleascii-sites.txt
6 index=0
7
8 cd $HOME
9 mkdir FDA_$DATE
10 cd FDA_$DATE
11
12 cat $SITES | while read line; do
13     let "index++"
14     pad=$(printf "%05d" $index)
15     mkdir $DATE-$pad
16     cd $DATE-$pad
17     /usr/bin/xvfb-run -a -s "-screen 0 1280x1024x24" /usr/bin/wkhtmltopdf --dpi 200 --page-size Letter --custom-header 'User-Agent: Heritrix' $line $DATE-$pad.pdf
18     /usr/bin/xvfb-run -a -s "-screen 0 1280x1024x24" /usr/local/bin/wkhtmltoimage --use-xserver --custom-header 'User-Agent: Heritrix' $line $DATE-$pad.png
19     rm tmp.png
20     /usr/local/bin/wget --adjust-extension --page-requisites --convert-links --no-parent --random-wait --warc-file=$DATE-$pad.warc $line
21     cd $HOME/FDA_$DATE
22     zip -r $DATE-$pad.zip $DATE-$pad
23     rm -rf $DATE-$pad
24     echo "$(date) - $line archived" >> /var/log/daleascii.log
25 done
```

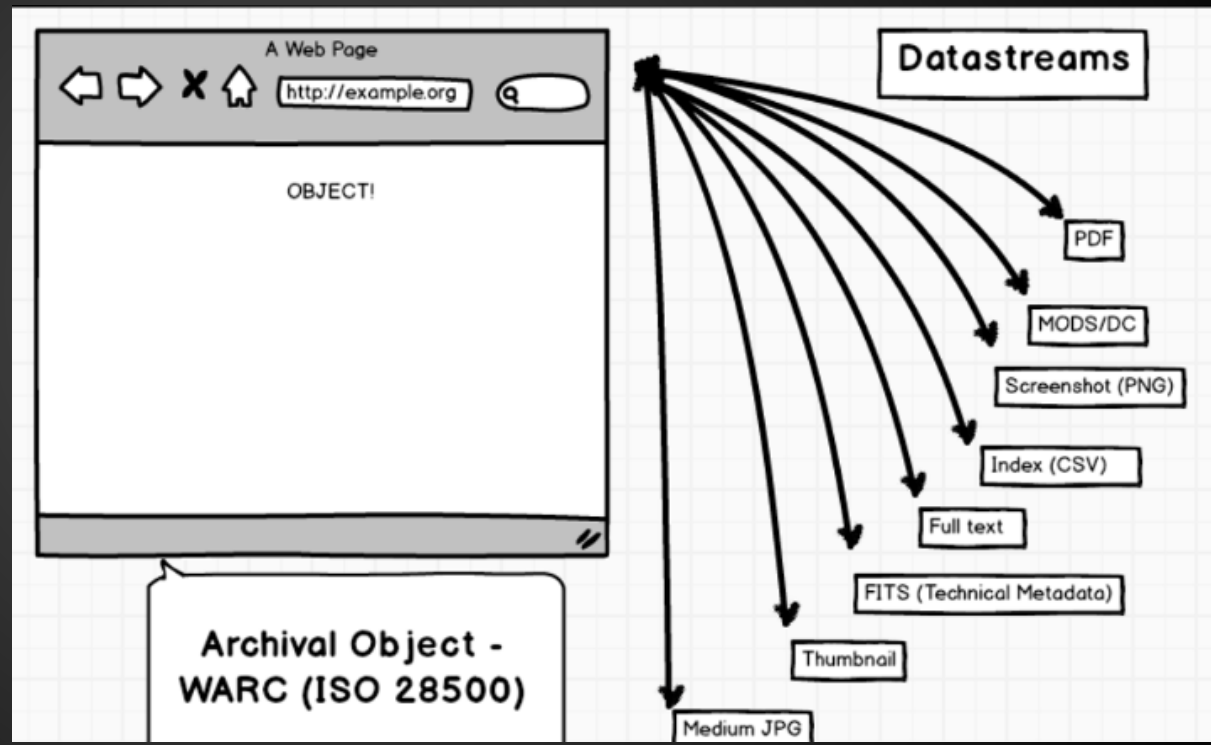
# AIP - Archival Information Package

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**

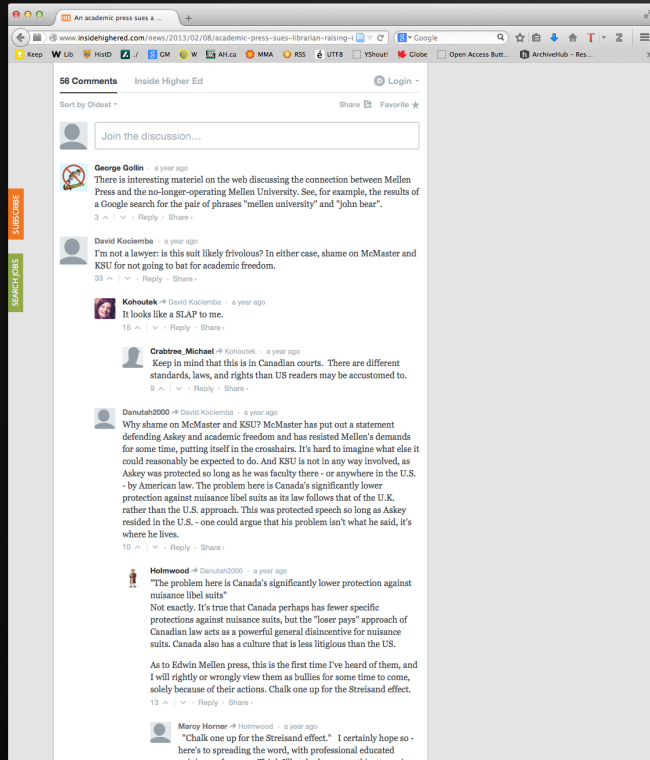
**An Ideal Case Study?**

# A Historian in the Archive

- \* WARC-Tools (<https://code.google.com/p/warc-tools/>)
- \* By Date (awesome!)

```
1. bash
Last login: Tue Jun 3 09:49:49 on console
Ians-MacBook-Pro:~ ianmilligan1$ cd desktop
Ians-MacBook-Pro:desktop ianmilligan1$ cd fda-by-date
Ians-MacBook-Pro:fda-by-date ianmilligan1$ ls
2013_02_20      2013_03_26      2013_06_24
2013_02_21      2013_03_27      2013_06_26
2013_02_22      2013_03_28      2013_06_28
2013_02_23      2013_03_29      2013_07_05
2013_02_24      2013_03_30      2013_07_06
2013_02_25      2013_04_02      2013_07_08
2013_02_26      2013_04_03      2013_07_12
2013_02_27      2013_04_04      2013_07_15
2013_02_28      2013_04_06      2013_07_16
2013_03_01      2013_04_09      2013_07_19
2013_03_02      2013_04_11      2013_07_21
2013_03_03      2013_04_12      2013_07_22
2013_03_04      2013_04_13      2013_07_26
2013_03_05      2013_04_16      2013_07_27
2013_03_06      2013_04_20      2013_07_28
2013_03_07      2013_04_22      2013_07_29
2013_03_08      2013_04_24      2013_08_01
2013_03_09      2013_04_25      2013_08_02
2013_03_10      2013_04_28      2013_08_06
2013_03_12      2013_05_06      2013_08_12
2013_03_13      2013_05_08      2013_08_14
2013_03_14      2013_05_15      all-by-date
2013_03_15      2013_05_17      all-by-date2
2013_03_16      2013_05_18      all-frequency
2013_03_17      2013_05_22      compile.sh
2013_03_18      2013_05_23      fda-topic-state-50.gz
2013_03_19      2013_05_24      fda_composition-50.txt
2013_03_20      2013_05_26      fda_keys-50.txt
2013_03_21      2013_05_28      ner
2013_03_22      2013_06_05      topics
2013_03_23      2013_06_09      wiki-editlist
2013_03_24      2013_06_18      wiki-scrape
2013_03_25      2013_06_22
Ians-MacBook-Pro:fda-by-date ianmilligan1$
```

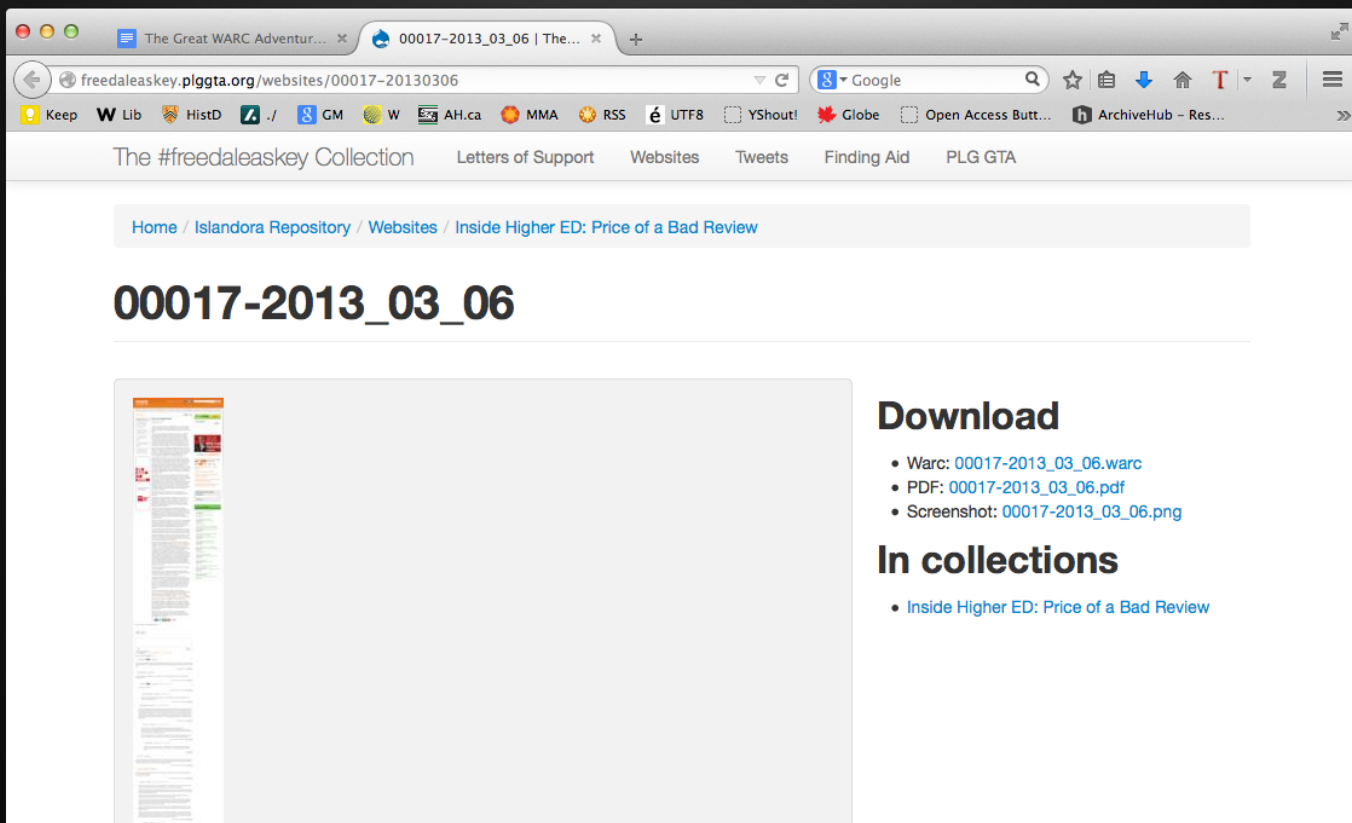
# A Historian in the Archive



- \* ~~Don't~~ read the comments
- \* Disqus



# A Historian in the Archive



The screenshot shows a web browser window with the address bar displaying `freedaleaskey.plgga.org/websites/00017-20130306`. The page title is "00017-2013\_03\_06 | The...". The browser's address bar also shows "Google" and various search and navigation icons. The page content includes a navigation menu with items like "The #freedaleaskey Collection", "Letters of Support", "Websites", "Tweets", "Finding Aid", and "PLG GTA". Below the navigation menu is a breadcrumb trail: "Home / Islandora Repository / Websites / Inside Higher ED: Price of a Bad Review". The main heading is "00017-2013\_03\_06". To the left of the main content is a vertical thumbnail image of a document page. To the right, under the heading "Download", there is a list of three items: "Warc: 00017-2013\_03\_06.warc", "PDF: 00017-2013\_03\_06.pdf", and "Screenshot: 00017-2013\_03\_06.png". Below this, under the heading "In collections", there is a single item: "Inside Higher ED: Price of a Bad Review".

The Great WARC Adventur... x 00017-2013\_03\_06 | The... x +

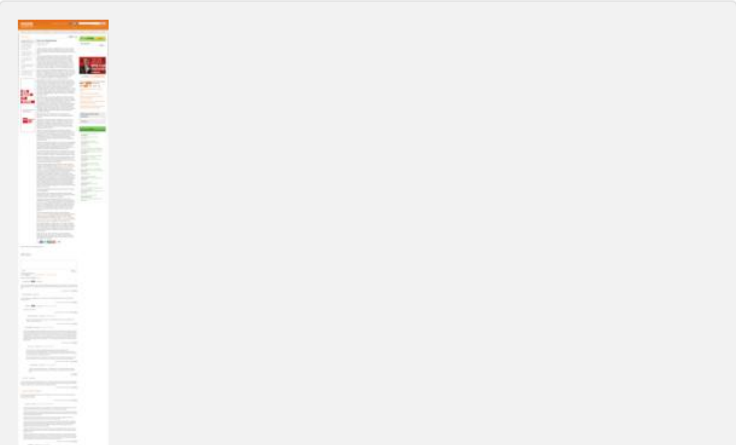
freedaleaskey.plgga.org/websites/00017-20130306

Keep W Lib HistD GM W AH.ca MMA RSS UTF8 YShout! Globe Open Access Butt... ArchiveHub - Res...

The #freedaleaskey Collection Letters of Support Websites Tweets Finding Aid PLG GTA

Home / Islandora Repository / Websites / Inside Higher ED: Price of a Bad Review

## 00017-2013\_03\_06



### Download

- Warc: [00017-2013\\_03\\_06.warc](#)
- PDF: [00017-2013\\_03\\_06.pdf](#)
- Screenshot: [00017-2013\\_03\\_06.png](#)

### In collections

- [Inside Higher ED: Price of a Bad Review](#)

# Learning from Word Frequency

Google Ngram Viewer

https://books.google.com/ngrams/graph?content=nationalize%2Cprivatize&year\_start=1900&year...

Keep W Lib HistD ./. GM W AH.ca MMA RSS UTF8 YShout! Globe Open Access Butt... ArchiveHub - Res... Most Visited >>

# Google books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)

[Share](#)  [Tweet](#) [Embed Chart](#)

Year	nationalize (%)	privatize (%)
1900	0.000100	0.000000
1910	0.000100	0.000000
1920	0.000180	0.000000
1930	0.000180	0.000000
1940	0.000230	0.000000
1950	0.000320	0.000000
1960	0.000380	0.000000
1970	0.000380	0.000000
1980	0.000350	0.000010
1990	0.000300	0.000150
2000	0.000200	0.000550

(click on line/label for focus)

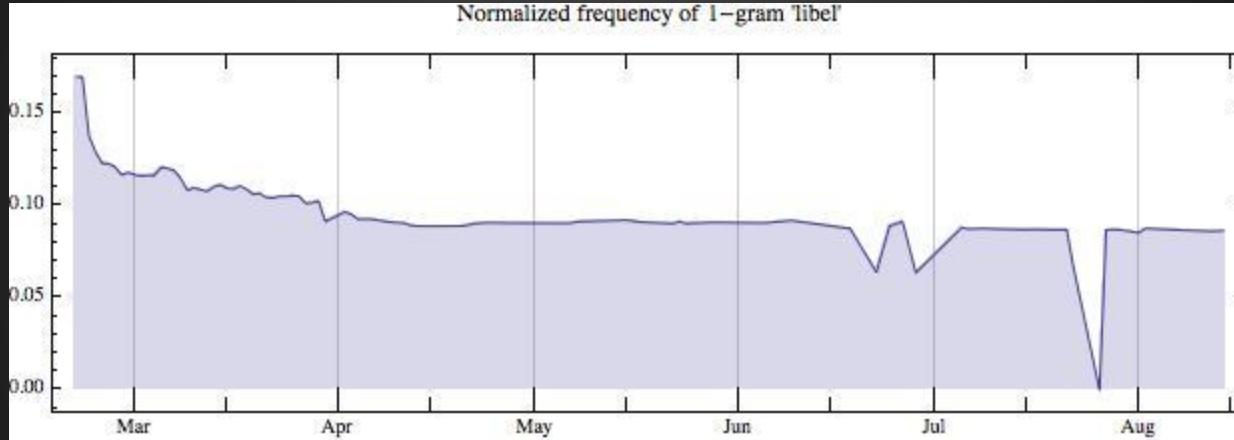
Search in Google Books:

<a href="#">1900 - 1920</a>	<a href="#">1921 - 1961</a>	<a href="#">1962 - 1967</a>	<a href="#">1968 - 1989</a>	<a href="#">1990 - 2000</a>	<a href="#">nationalize</a>	English
<a href="#">1900 - 1987</a>	<a href="#">1988 - 1996</a>	<a href="#">1997</a>	<a href="#">1998</a>	<a href="#">1999 - 2000</a>	<a href="#">privatize</a>	English

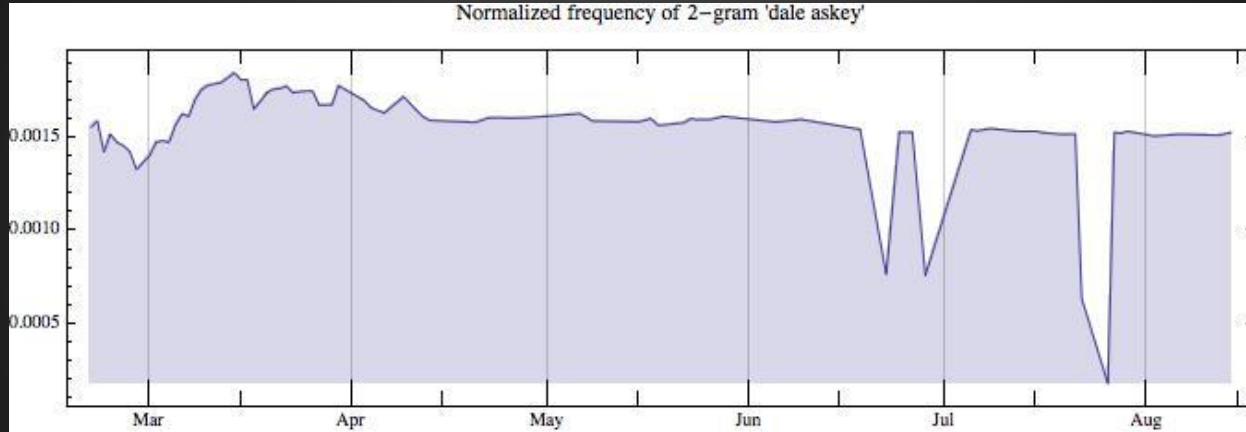
Run your own experiment! Raw data is available for download [here](#).

© 2013 Google - [Privacy & Terms](#) - [About Google](#) - [About Google Books](#) - [About Ngram Viewer](#)

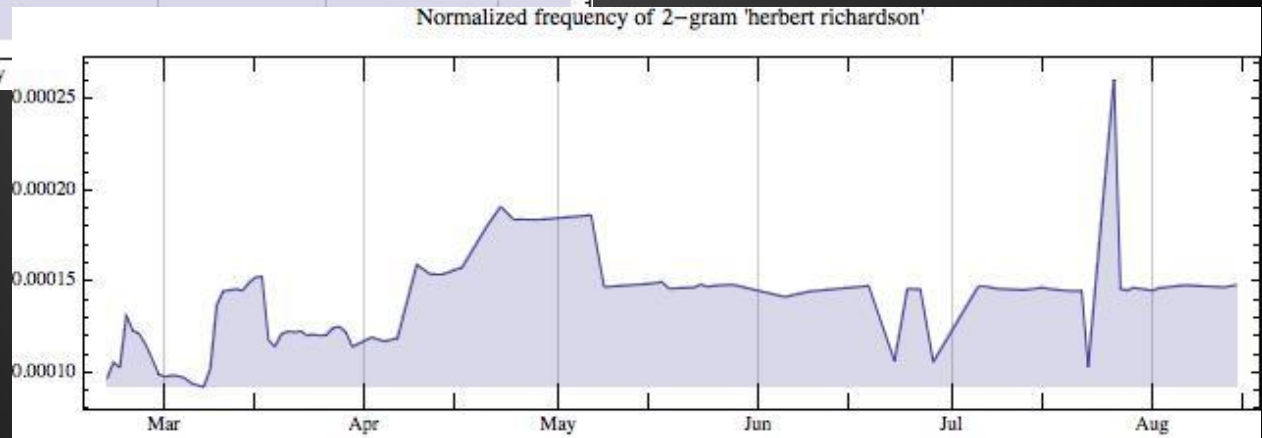
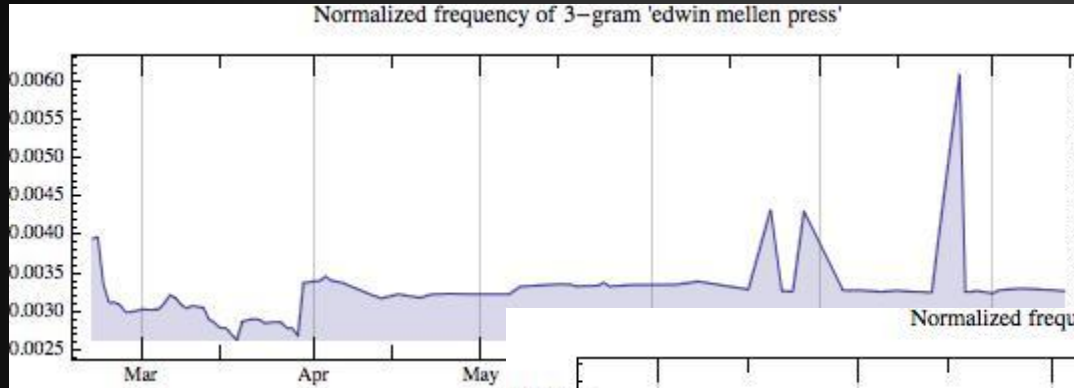
# A Historian in the Archive (distant reading)



# 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



The Globe an...



National Post



Toronto Sun



Yahoo News ...



Canada.com

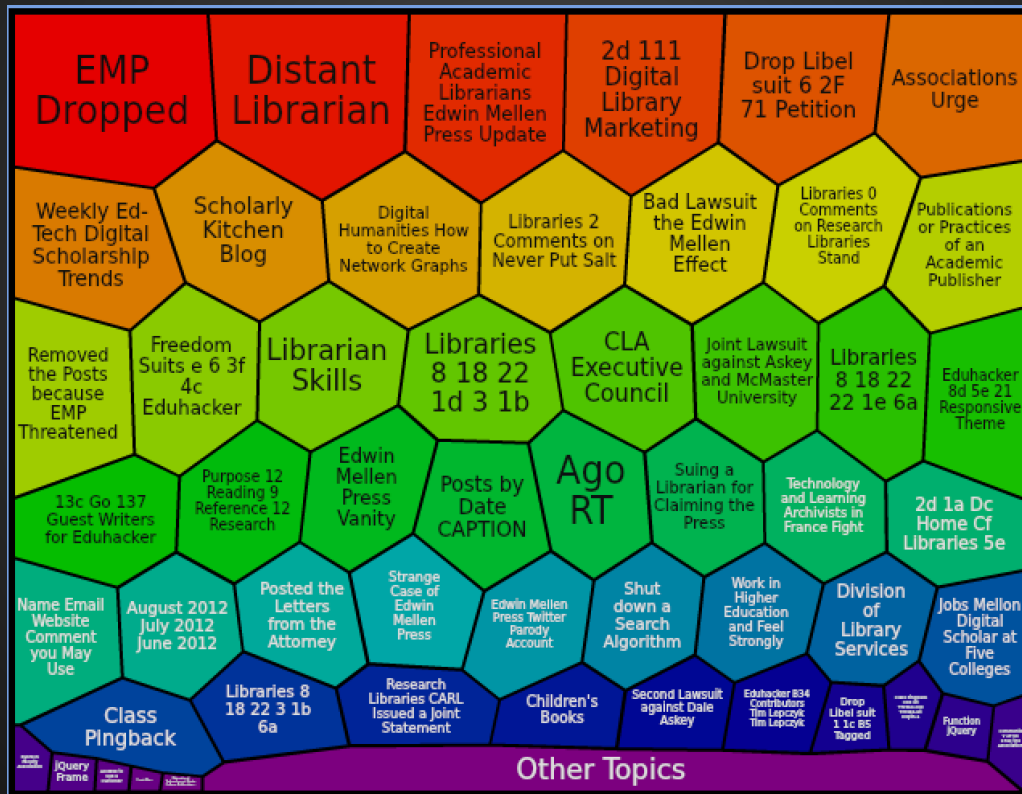


CBC.ca

**Explore in depth** (638 more articles)



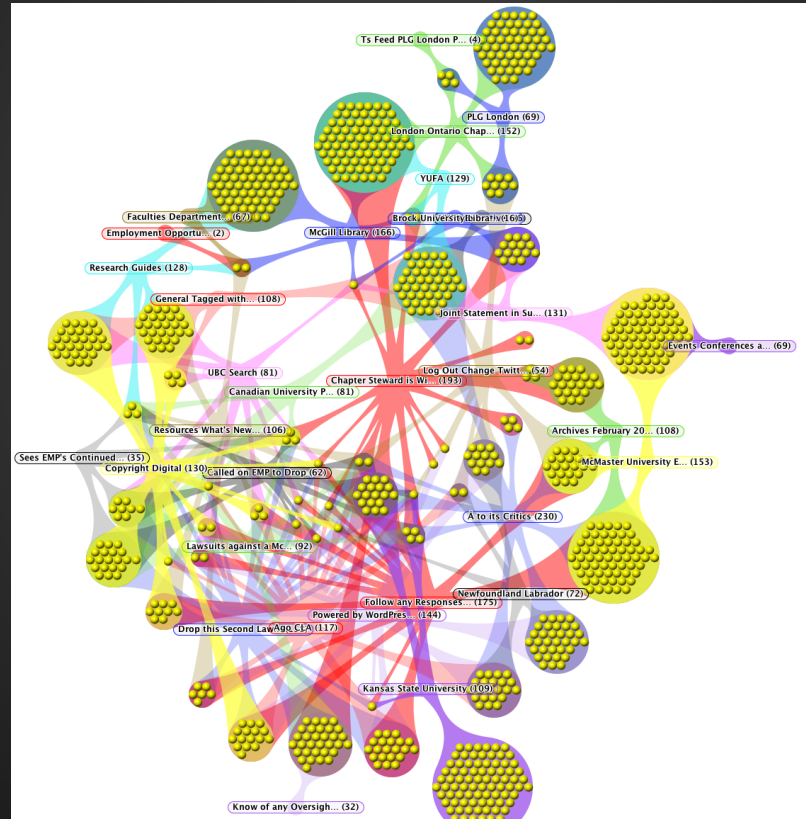
# A Historian in the Archive (distant reading) - search 'edwin mellen'



# 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)

31	1191	http chronicle ww techdirt eduhacker ca edwin mellen net ala scholarly brocku linguafranca bibvirdev digital uqac html msg post	
1	23	http category libraries blogs ca blog librarian dal https ww academic index wordpress infodocket google freedom law org social	
11	1366	http ww sspnet scholarlykitchen librarian html gn php ca press twitter posts wordpress tag action chronicle libraries wiredcampus oif	
2	734	http ww org ca html utlibrarians category press net edwin library lawsuit libraryjournal wordpress blog lj content public librarian	
32	3991	http ww reuters insidehighered wgrz buffalonews leiterreports blogs utlibrarians media news article apps university slashdot dll pbc default wordpress	
14	0992	http mellen utlibrarians net wordpress ca edwin category student ww publishing blogspot dale letters lawsuits var threatens libraries folders	
43	4036	http ca net blogspot ww library edwin search wordpress public html var email academic label comments press tag law	
12	8179	http ww ubc html kcoyle ca de index library bideutschland deutsch mr aktuelles press wordpress academic infodocket post arl	
5	831	http ww mellen suit university ca blog html librarians ala events drops mcmaster void stephanielgrossmlis canadian news science march	
4	616	blogspot label outofthejungle search http ca html blog ww wordpress javascript void askey org net gn librarians libraries mellen	
33	5445	http ca ww javascript blogs library press feed legal law reviews php gavialib mellen july category typepad hours utf	

# A Historian in the Archive (distant reading)

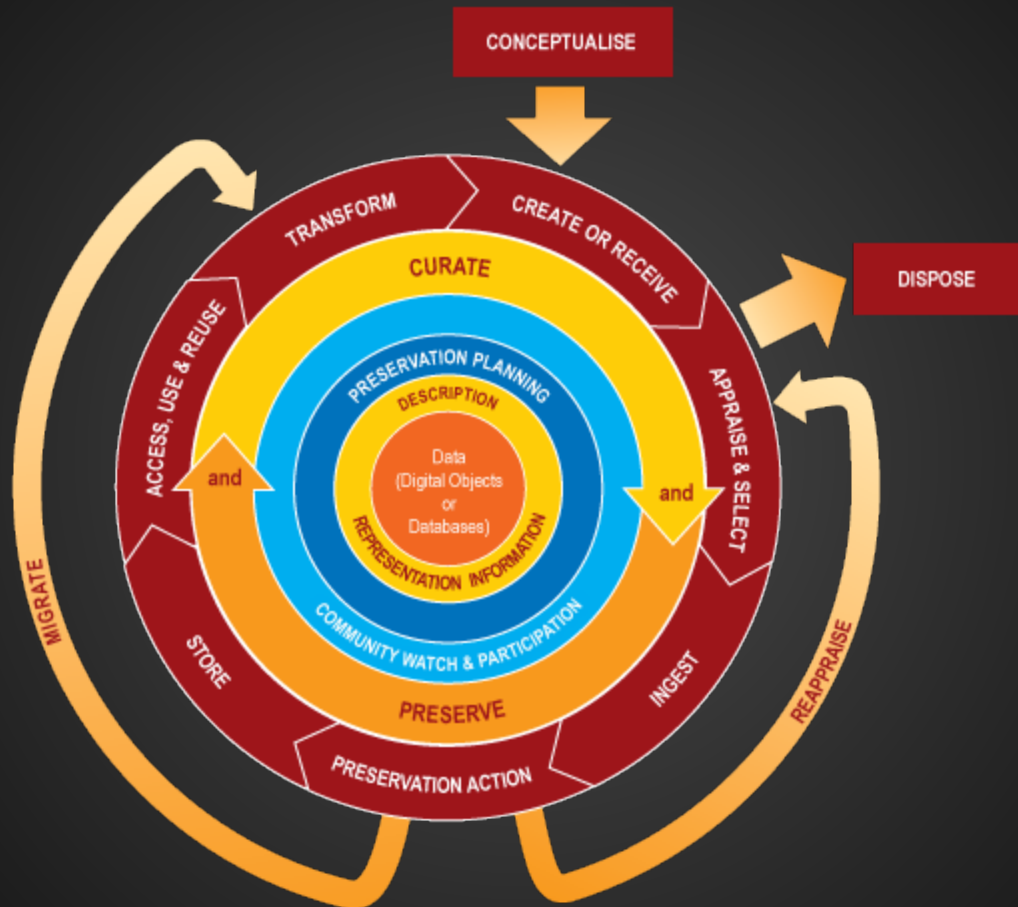
7	997	http ww librarian blogs libraryjournal university mellen lawsuit suing page wordpress php content tag askey author mcmaster ca sues	
13	5084	ww http wordpress net askey html freedom libraries tag reddit yzlj scholarlykitchen research press de sspnet javascript jl folders	
22	0276	http library mellen tag ww category job kvps university academic blog blogs tmp librarian scholarship https text feed review	
47	3865	ww http edwin post wordpress ca libraryjournal books lj february category press sued article utlibrarians chronicle philosophy feed dale	
37	0637	http ww ca cbc html void utlibrarians college wordpress academic libraries category comments million net quality mcmaster critical tmp	
45	0932	http org ww ca facebook university comment dale askey php libraries mcmaster user library slaw academic book var service	
15	3549	http html blogspot ca ww net org lj university utlibrarians mlanet category libraries publisher javascript news olaweb google email	
9	621	http org slashdot ww file news ca kvps html sid pl comments folders jl saycampuslife gn yzlj var cid	
42	3097	http ww ca html blog february askey news wordpress free references utfa view kvps speech connection intellectual category press	
29	5192	http ca ww org html university content blogger comments posts net canada url librarian edwin publisher association mcmaster professional	

**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](mailto:ruestn@yorku.ca) / @ruebot

Ian Milligan: [i2milligan@uwaterloo.ca](mailto:i2milligan@uwaterloo.ca) / @ianmilligan1