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

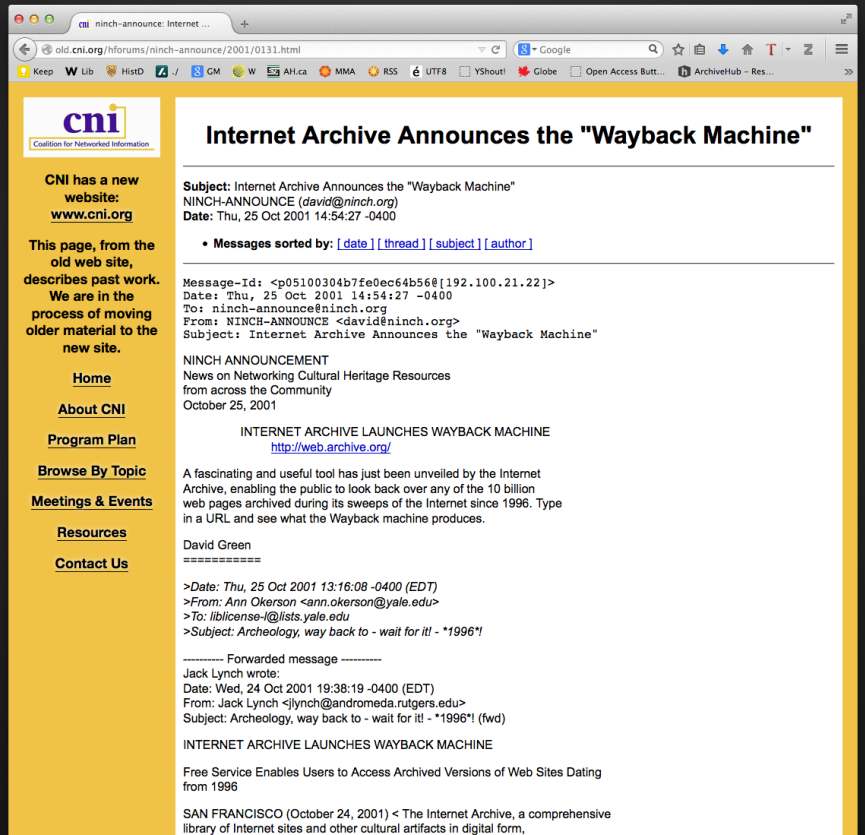
Historical overview of web archiving

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

The Internet Archive



ninch-announce: Internet ...

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

cni
Coalition for Networked Information

CNI has a new website: 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*)
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 Networking 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*!*

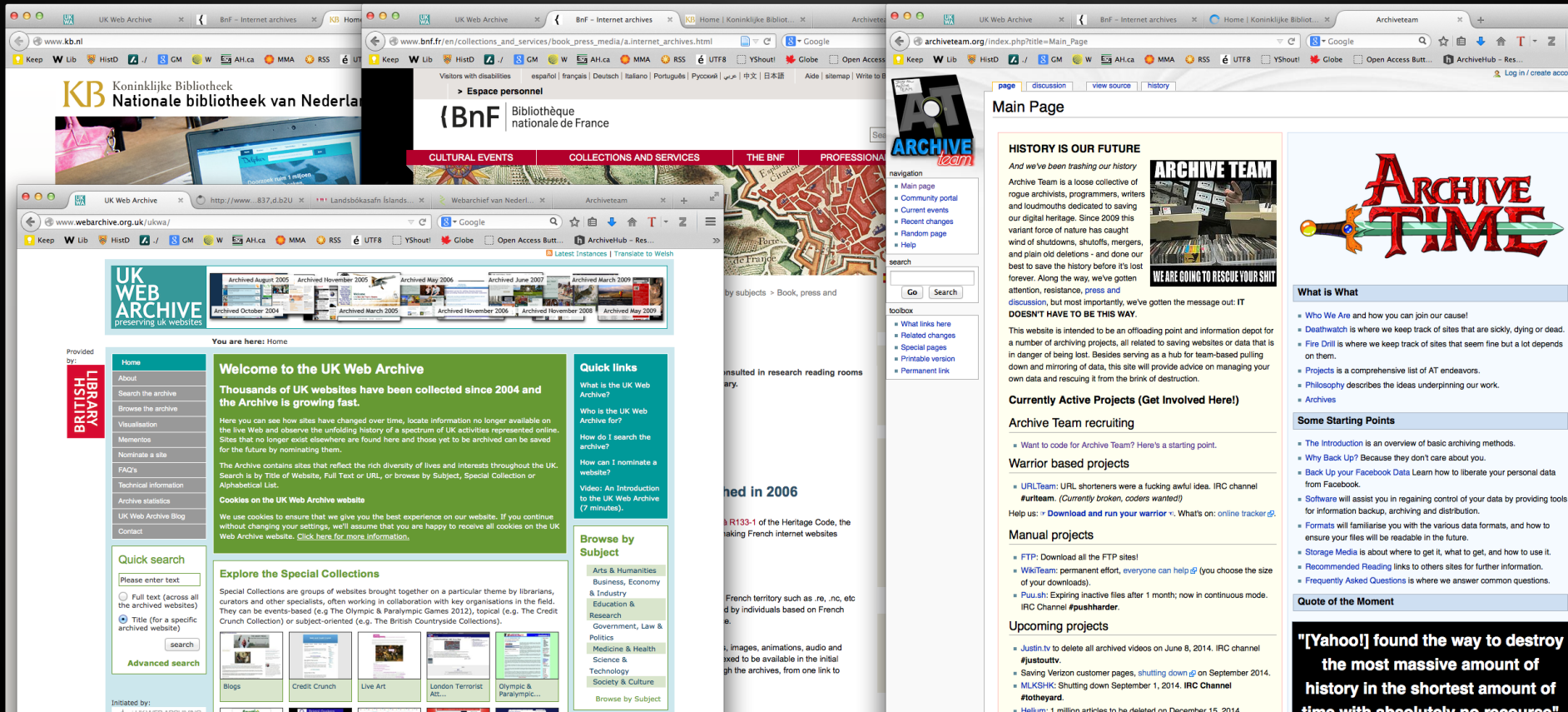
--------- Forwarded message ---------
Jack Lynch wrote:
Date: Wed, 24 Oct 2001 19:38:19 -0400 (EDT)
From: Jack Lynch <jlynch@andromeda.rutgers.edu>
Subject: Archeology, way back to - wait for it! - *1996*! (fwd)

INTERNET ARCHIVE LAUNCHES WAYBACK MACHINE

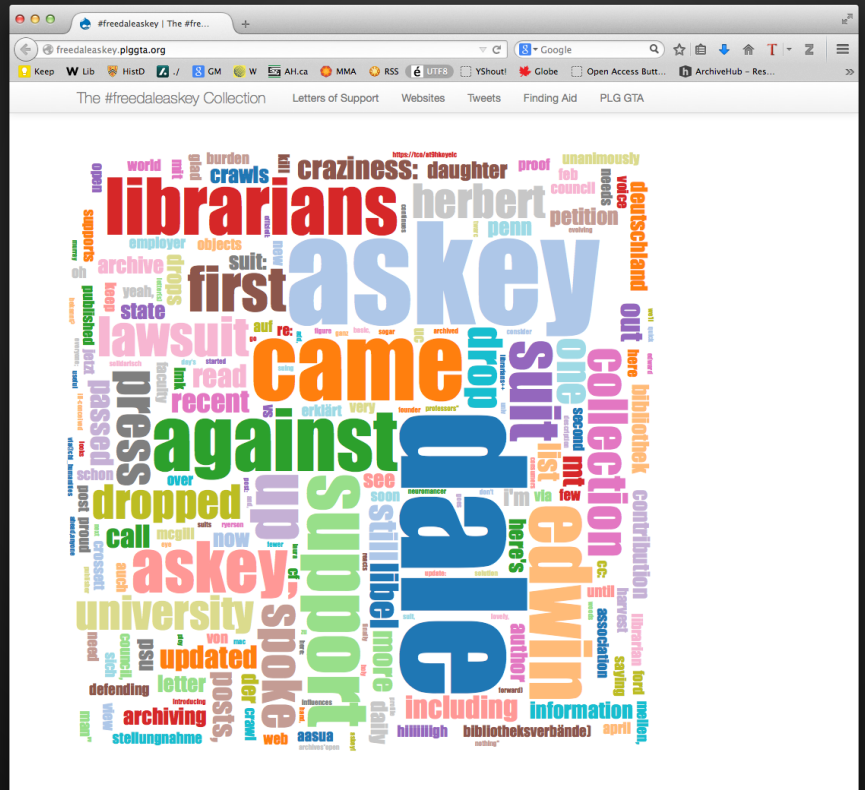
Free Service Enables Users to Access Archived Versions of Web Sites Dating from 1996

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

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 8, 2013 by plggta

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.

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'

     0K ..                                                    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:RMV6QLTEU6ECCRKM6SB6GLJUPBPQVFBN
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:F7GN2MHQRZ35EAZEUVYOVIXSJBCHQD7U
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>
```

https://digital.library.yorku.ca/yul-92640/biafra

PROVENANCE!

Harvest of BoycottMellenPress.com site.warc (~/Downloads) - VIM

```
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:TP3MEJQ7DVS36JOULUEE3DINUMKC36II^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
^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:QR775IO5R263VLKHSLCKM5QQPHW4RE2G^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
^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"^M
```

1,1 Top

Harvest of BoycottMellenPress.com site.warc (~/Downloads) - VIM

```
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>
@
@
@
                                                                    56,7          42%
```

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.jira.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

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

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

Find a Collection by Name | Search | *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. Iincludes 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

Find an Organization by Name | Search | *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

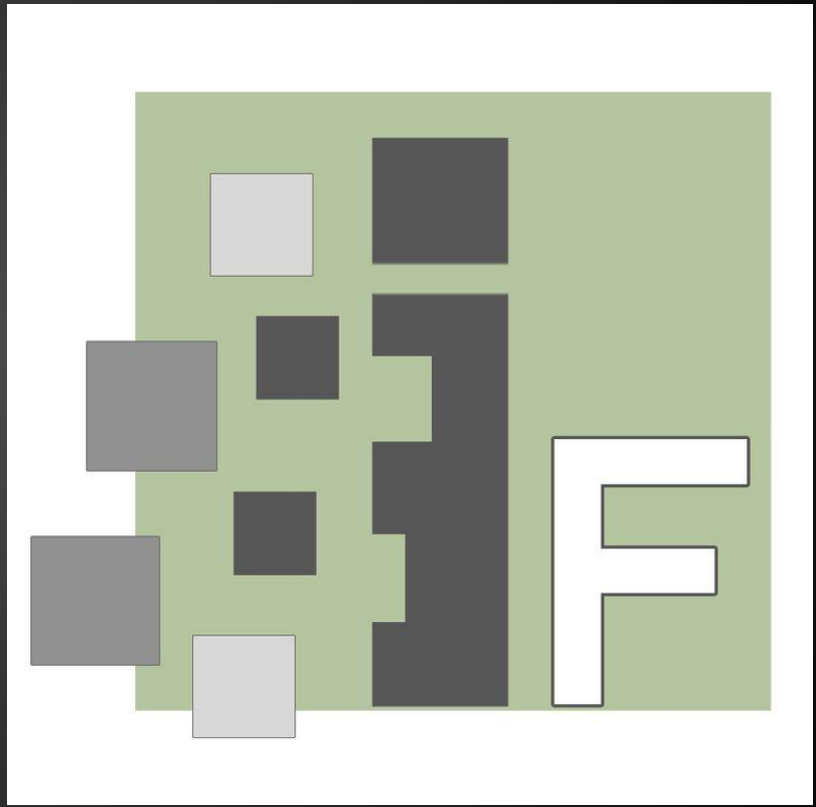
Columbia College Chicago is an international leader and recognized pioneer in arts and media education. With over 120 years of deep experience teaching creative students to…

$$$

Can we do this with Open Source software?

Use the tools you know

Drupal
Fedora Commons
Islandora
Heritrix
wkhtmltopdf
wkhtmltoimage

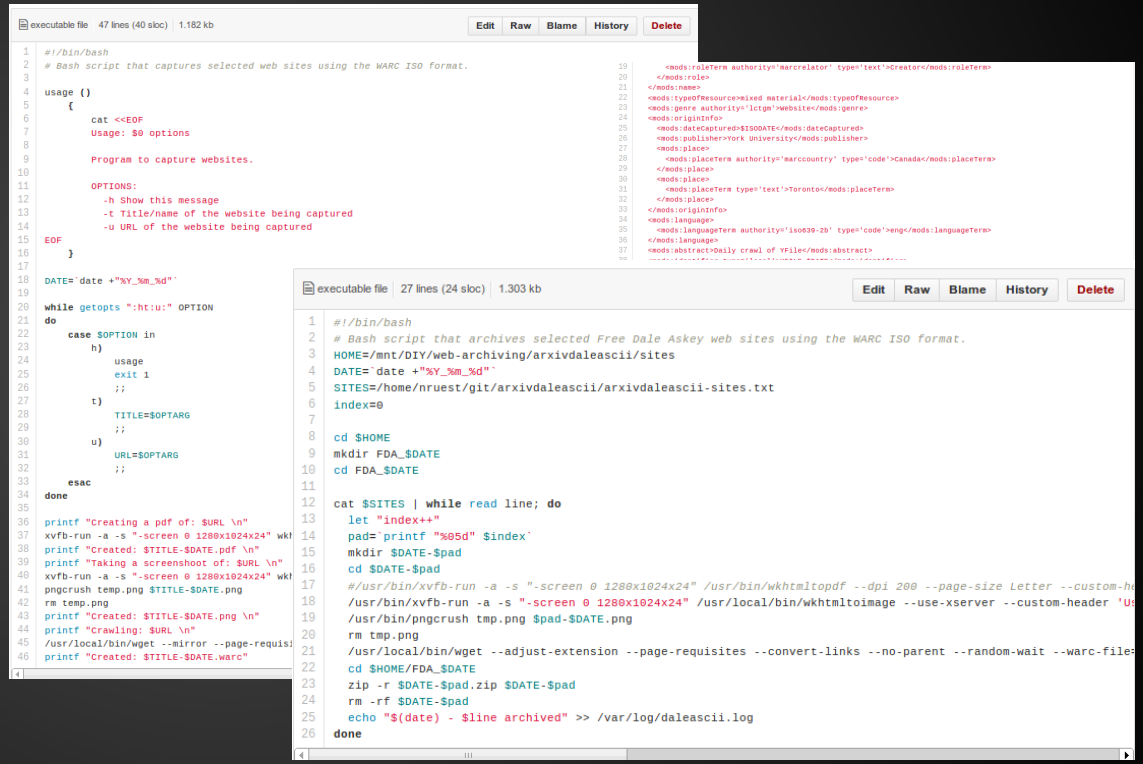


Can we talk OAIS?

SIP - Submission Information Package

Heritrix
Wget
wkhtmltopdf
wkhtmltoimage

...and Bash!



executable file | 47 lines (40 sloc) | 1.182 kb — Edit | Raw | Blame | History | Delete

```
#!/bin/bash
# Bash script that captures selected web sites using the WARC ISO format.

usage ()
    {
        cat <<EOF
        Usage: $0 options

        Program to capture websites.

        OPTIONS:
          -h Show this message
          -t Title/name of the website being captured
          -u URL of the website being captured
EOF
    }

DATE=`date +"%Y_%m_%d"`

while getopts ":ht:u:" OPTION
do
    case $OPTION in
        h)
            usage
            exit 1
            ;;
        t)
            TITLE=$OPTARG
            ;;
        u)
            URL=$OPTARG
            ;;
    esac
done

printf "Creating a pdf of: $URL \n"
xvfb-run -a -s "-screen 0 1280x1024x24" wkh
printf "Created: $TITLE-$DATE.pdf \n"
printf "Taking a screenshoot of: $URL \n"
xvfb-run -a -s "-screen 0 1280x1024x24" wkh
pngcrush temp.png $TITLE-$DATE.png
rm temp.png
printf "Created: $TITLE-$DATE.png \n"
printf "Crawling: $URL \n"
/usr/local/bin/wget --mirror --page-requis
printf "Created: $TITLE-$DATE.warc"
```

```
    <mods:roleTerm authority='marcrelator' type='text'>Creator</mods:roleTerm>
  </mods:role>
</mods:name>
<mods:typeOfResource>mixed material</mods:typeOfResource>
<mods:genre authority='lctgm'>Website</mods:genre>
<mods:originInfo>
  <mods:dateCaptured>$ISODATE</mods:dateCaptured>
  <mods:publisher>York University</mods:publisher>
  <mods:place>
    <mods:placeTerm authority='marccountry' type='code'>Canada</mods:placeTerm>
  </mods:place>
  <mods:place>
    <mods:placeTerm type='text'>Toronto</mods:placeTerm>
  </mods:place>
</mods:originInfo>
<mods:language>
  <mods:languageTerm authority='iso639-2b' type='code'>eng</mods:languageTerm>
</mods:language>
<mods:abstract>Daily crawl of YFile</mods:abstract>
```

executable file | 27 lines (24 sloc) | 1.303 kb — Edit | Raw | Blame | History | Delete

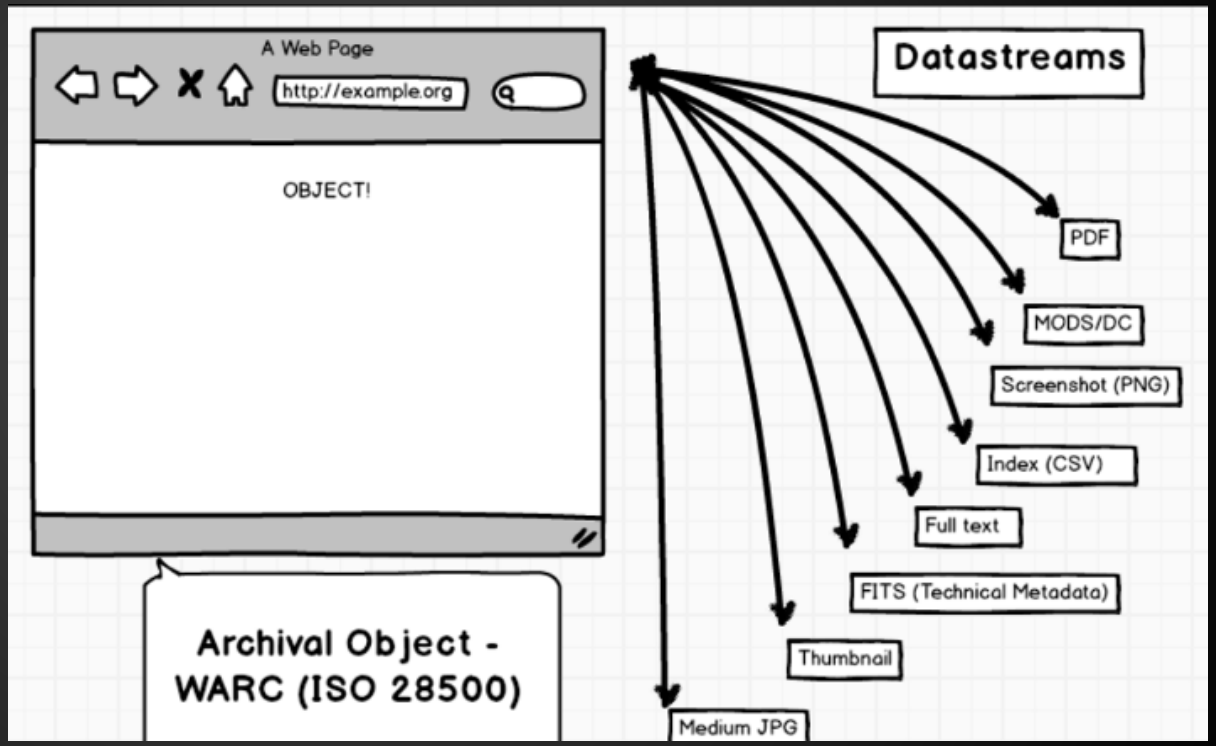
```
#!/bin/bash
# Bash script that archives selected Free Dale Askey web sites using the WARC ISO format.
HOME=/mnt/DIY/web-archiving/arxivdaleascii/sites
DATE=`date +"%Y_%m_%d"`
SITES=/home/nruest/git/arxivdaleascii/arxivdaleascii-sites.txt
index=0

cd $HOME
mkdir FDA_$DATE
cd FDA_$DATE

cat $SITES | while read line; do
  let "index++"
  pad=`printf "%05d" $index`
  mkdir $DATE-$pad
  cd $DATE-$pad
  #/usr/bin/xvfb-run -a -s "-screen 0 1280x1024x24" /usr/bin/wkhtmltopdf --dpi 200 --page-size Letter --custom-he
  /usr/bin/xvfb-run -a -s "-screen 0 1280x1024x24" /usr/local/bin/wkhtmltoimage --use-xserver --custom-header 'Us
  /usr/bin/pngcrush tmp.png $pad-$DATE.png
  rm tmp.png
  /usr/local/bin/wget --adjust-extension --page-requisites --convert-links --no-parent --random-wait --warc-file=
  cd $HOME/FDA_$DATE
  zip -r $DATE-$pad.zip $DATE-$pad
  rm -rf $DATE-$pad
  echo "$(date) - $line archived" >> /var/log/daleascii.log
done
```

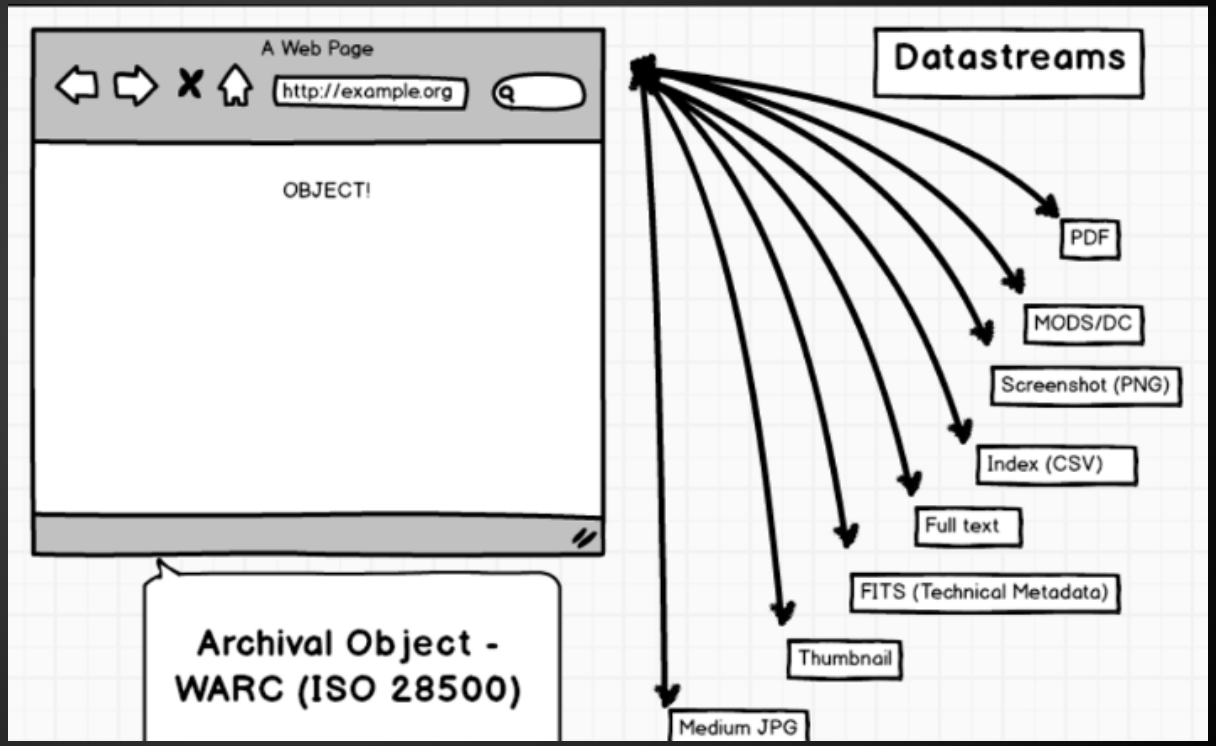
AIP - Archival Information Package

Drupal
Islandora
Fedora Commons
Web ARChive SP
Checksum
Checksum checker
Premis
FITS



DIP - Dissemination Information Package

Drupal
Islandora
Fedora Commons
Web ARChive SP

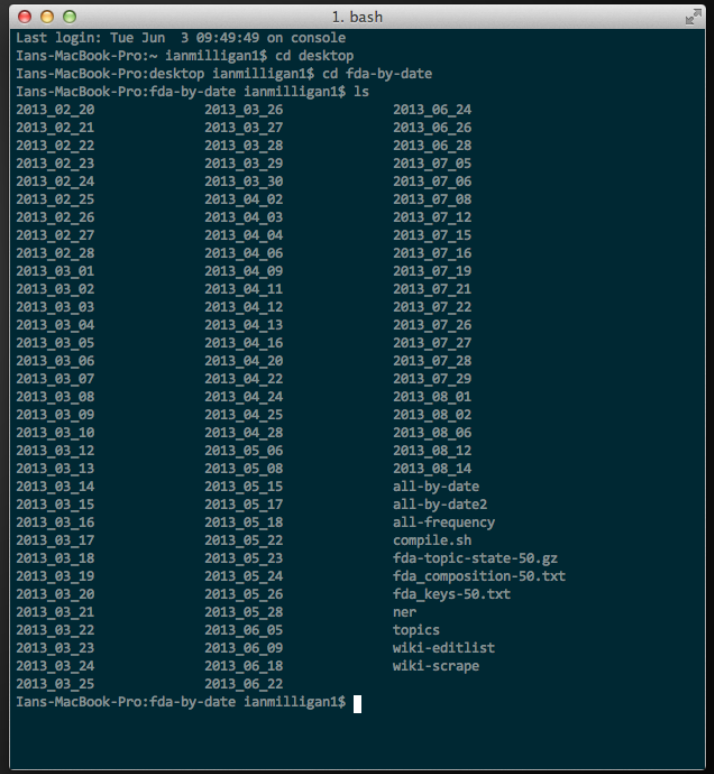


Interplay of the archivist and historian

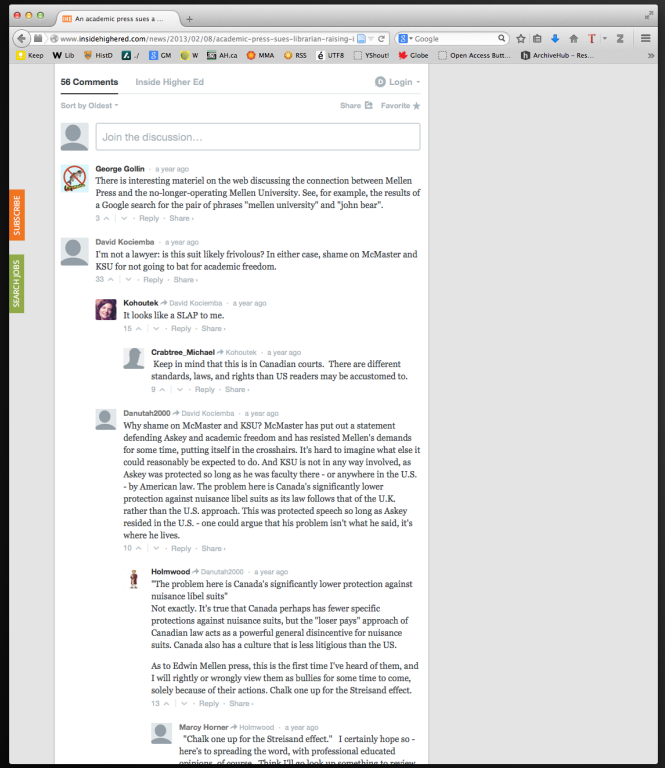
An Ideal Case Study?

A Historian in the Archive

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

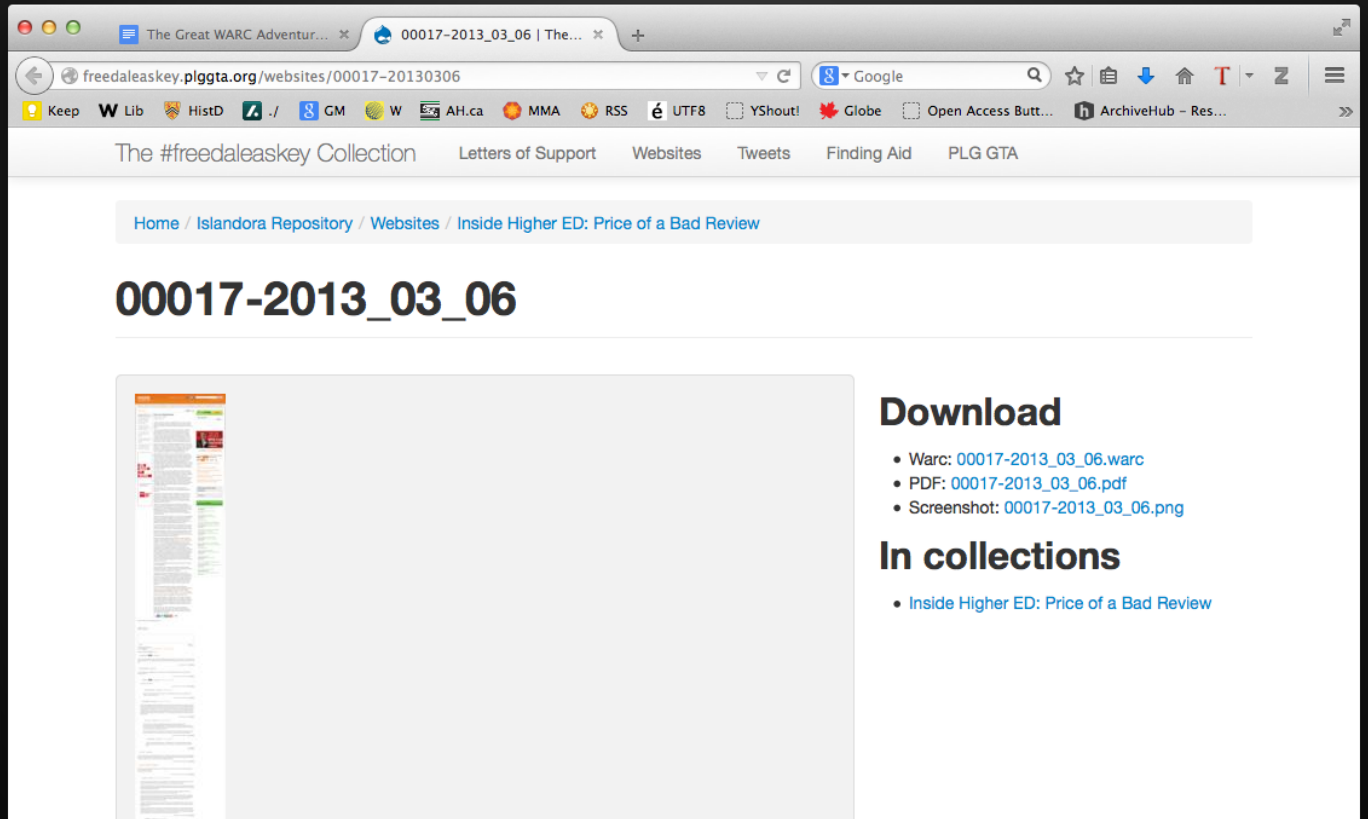


A Historian in the Archive



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

A Historian in the Archive



Learning from Word Frequency

Google Ngram Viewer

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

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

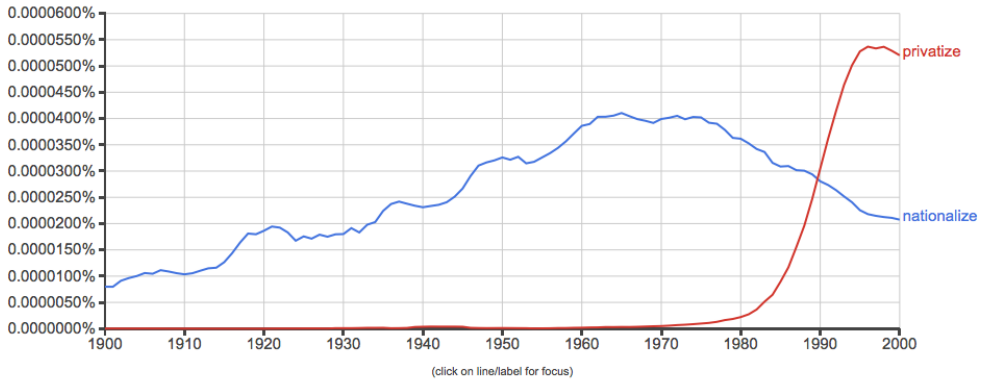
Google books Ngram Viewer

Graph these comma-separated phrases: nationalize,privatize case-insensitive

between 1900 and 2000 from the corpus English with smoothing of 3. Search lots of books

Share 0 Tweet

Embed Chart



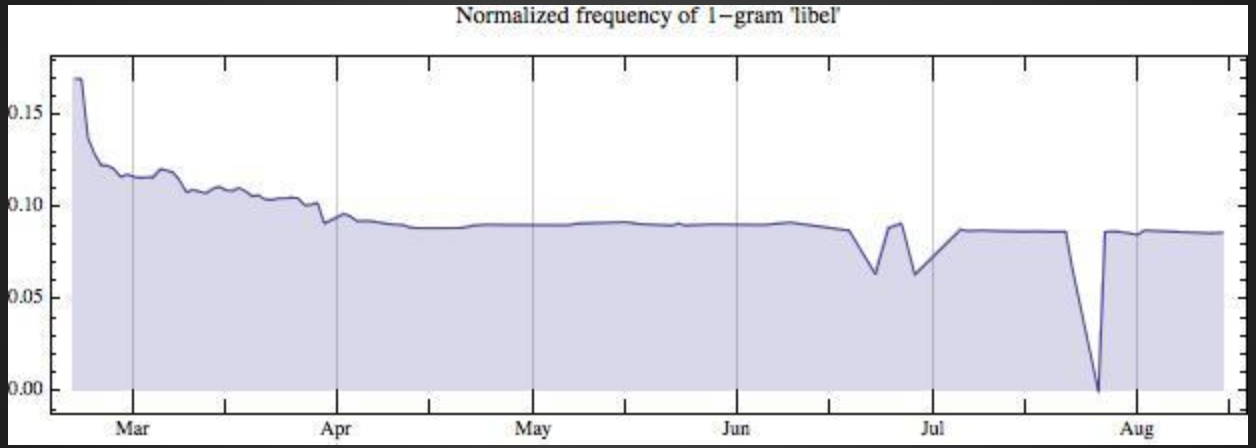
(click on line/label for focus)

Search in Google Books:

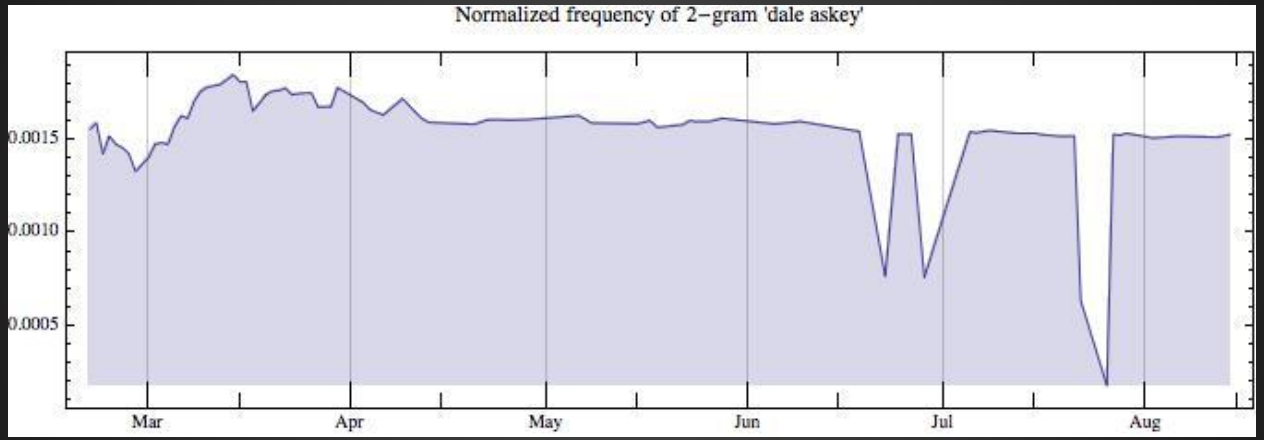
1900 - 1920	1921 - 1961	1962 - 1967	1968 - 1989	1990 - 2000	nationalize	English
1900 - 1987	1988 - 1996	1997	1998	1999 - 2000	privatize	English

Run your own experiment! Raw data is available for download here.

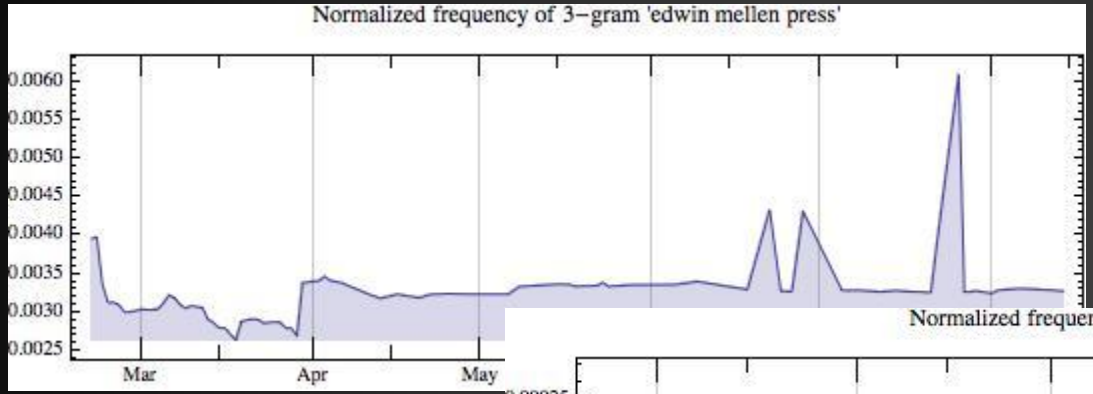
A Historian in the Archive (distant reading)

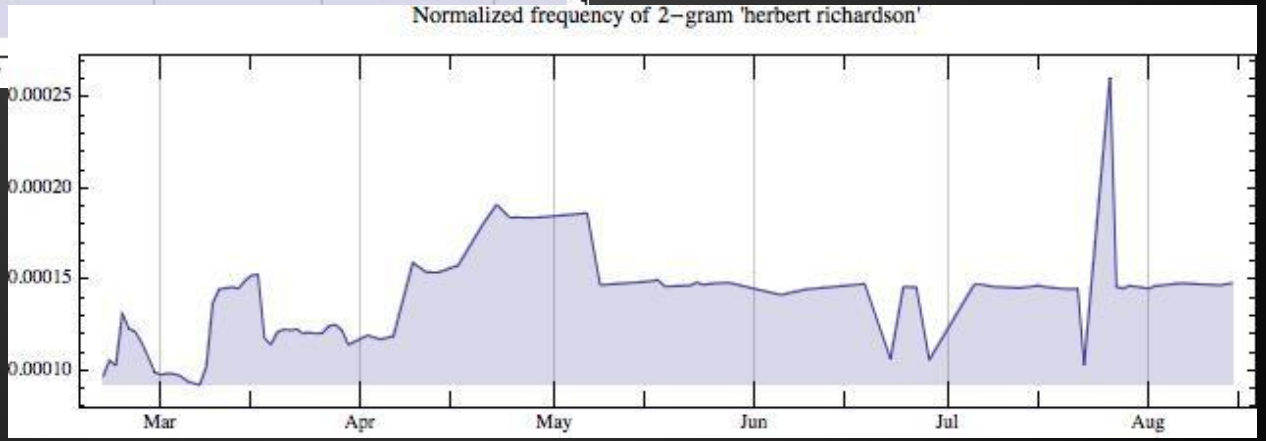


A Historian in the Archive (distant reading)



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

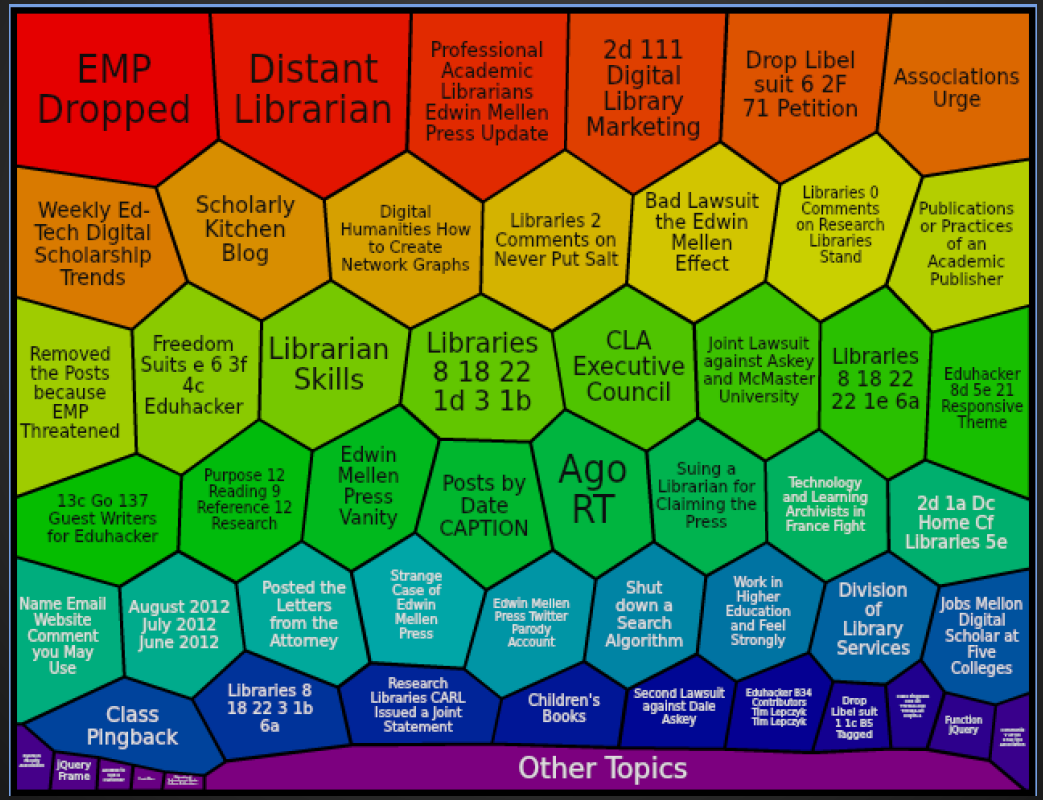

Yahoo News ...


Canada.com


CBC.ca

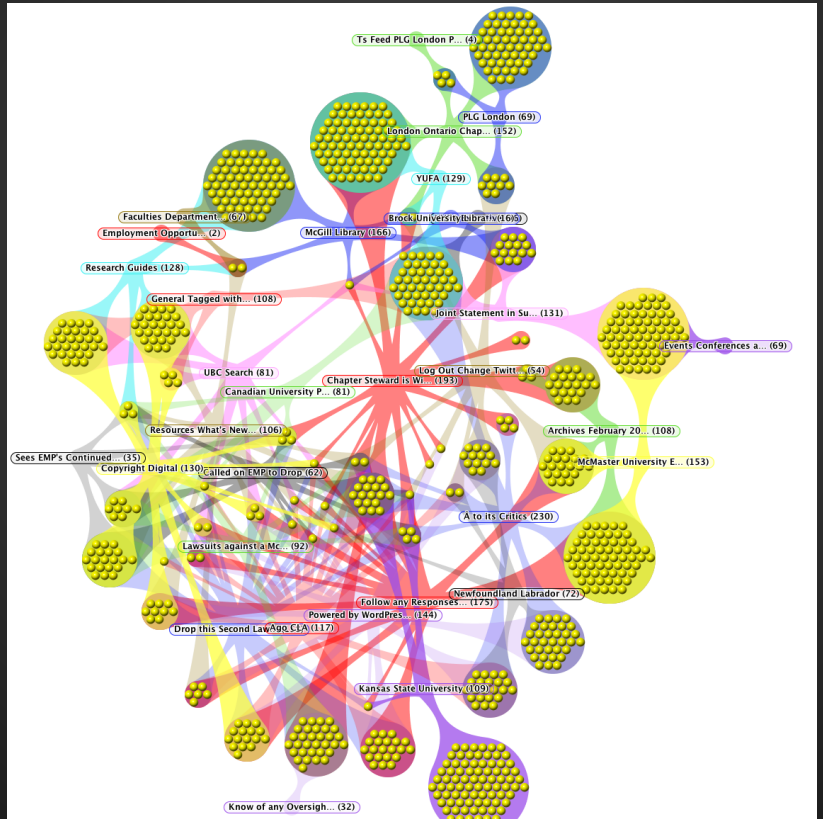
Explore in depth (638 more articles)

A Historian in the Archive (distant reading) - search ‘edwin mellen’



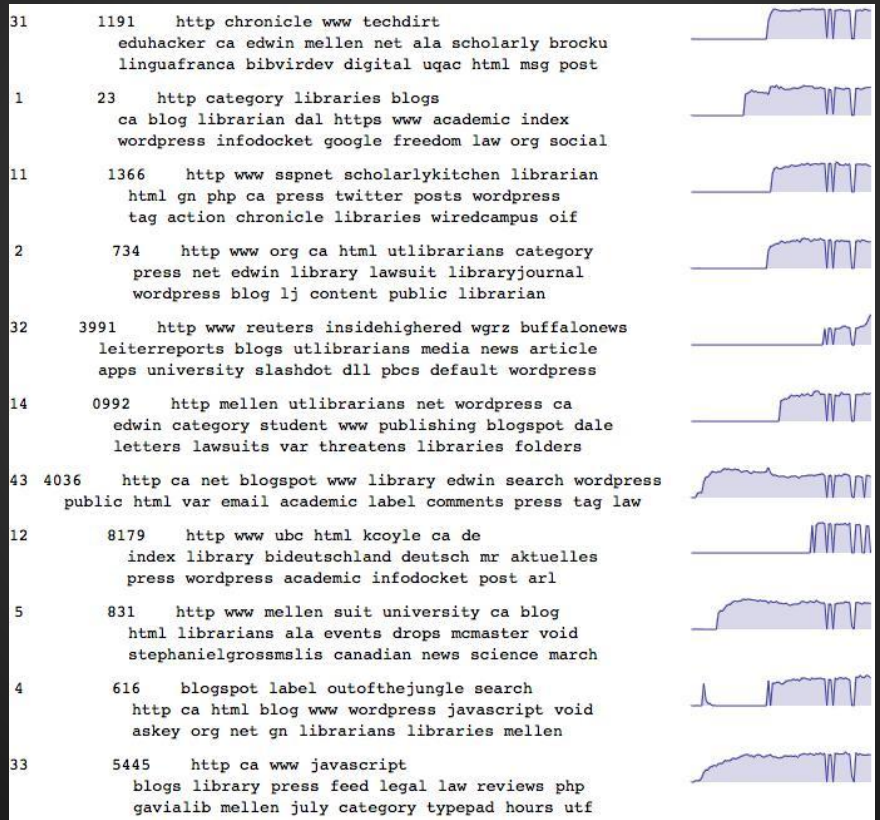
A Historian in the Archive (distant reading) - search ‘librarians’

A Historian in the Archive (distant reading) - search ‘dale askey’

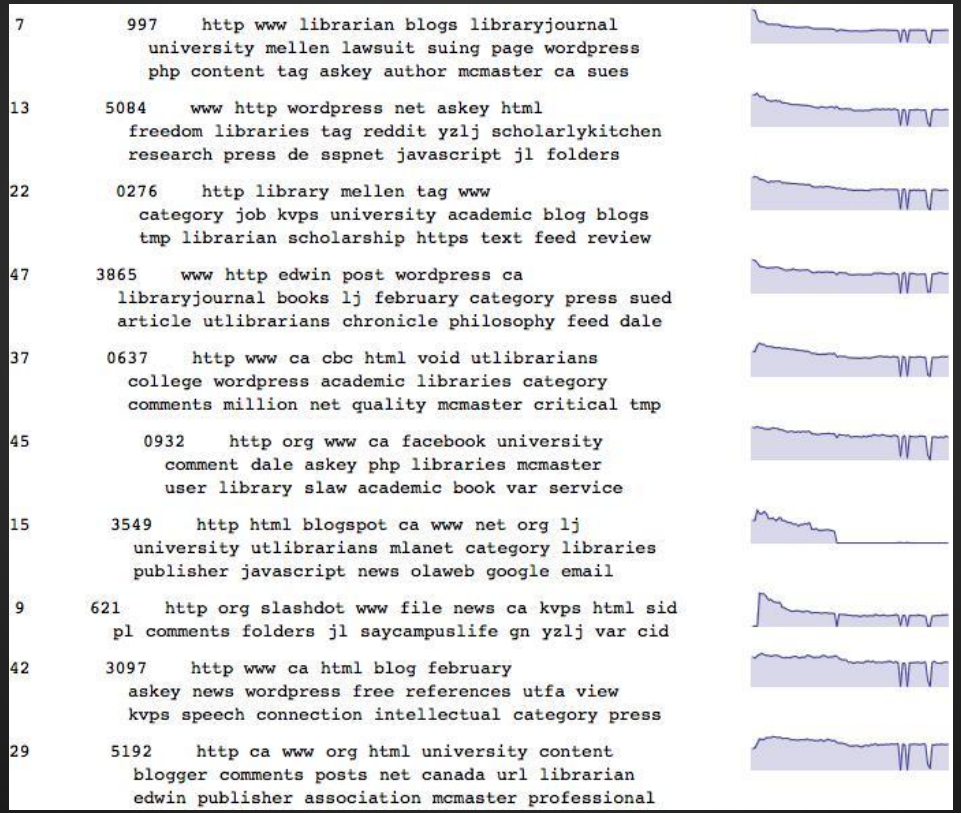


Problem: still need to know what you’re looking for!

A Historian in the Archive (distant reading)



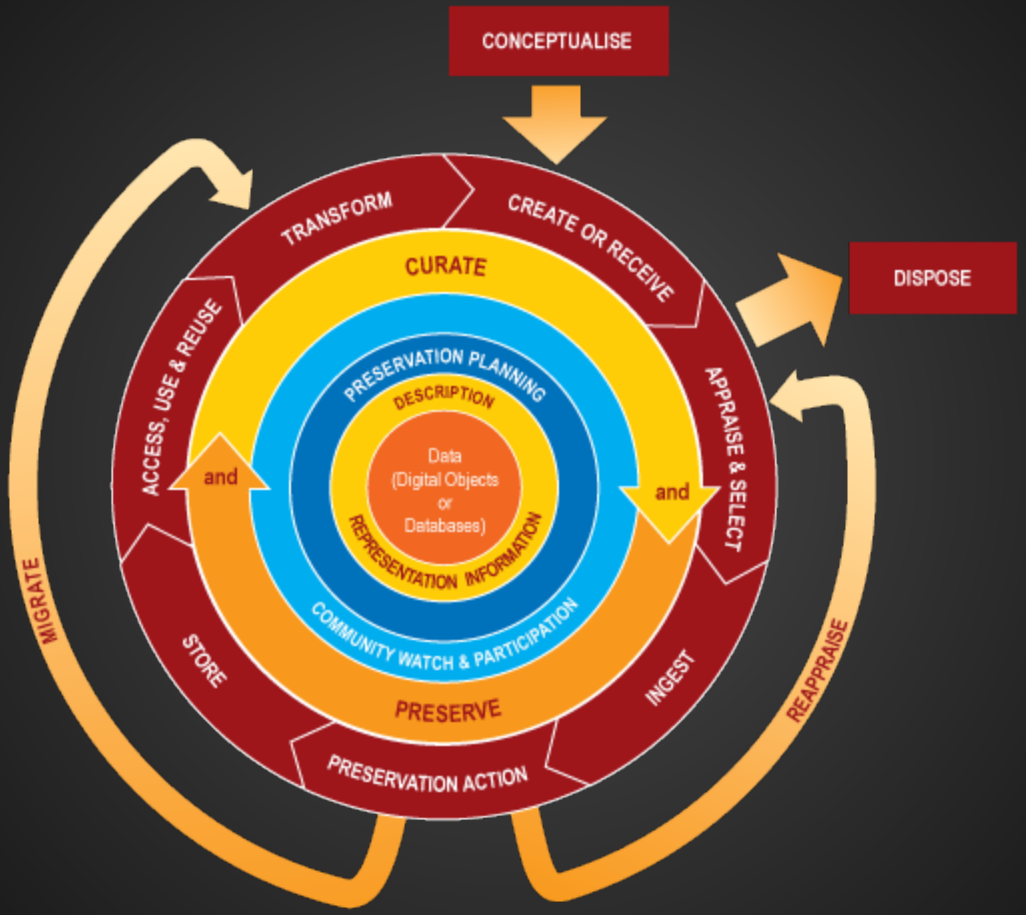
A Historian in the Archive (distant reading)

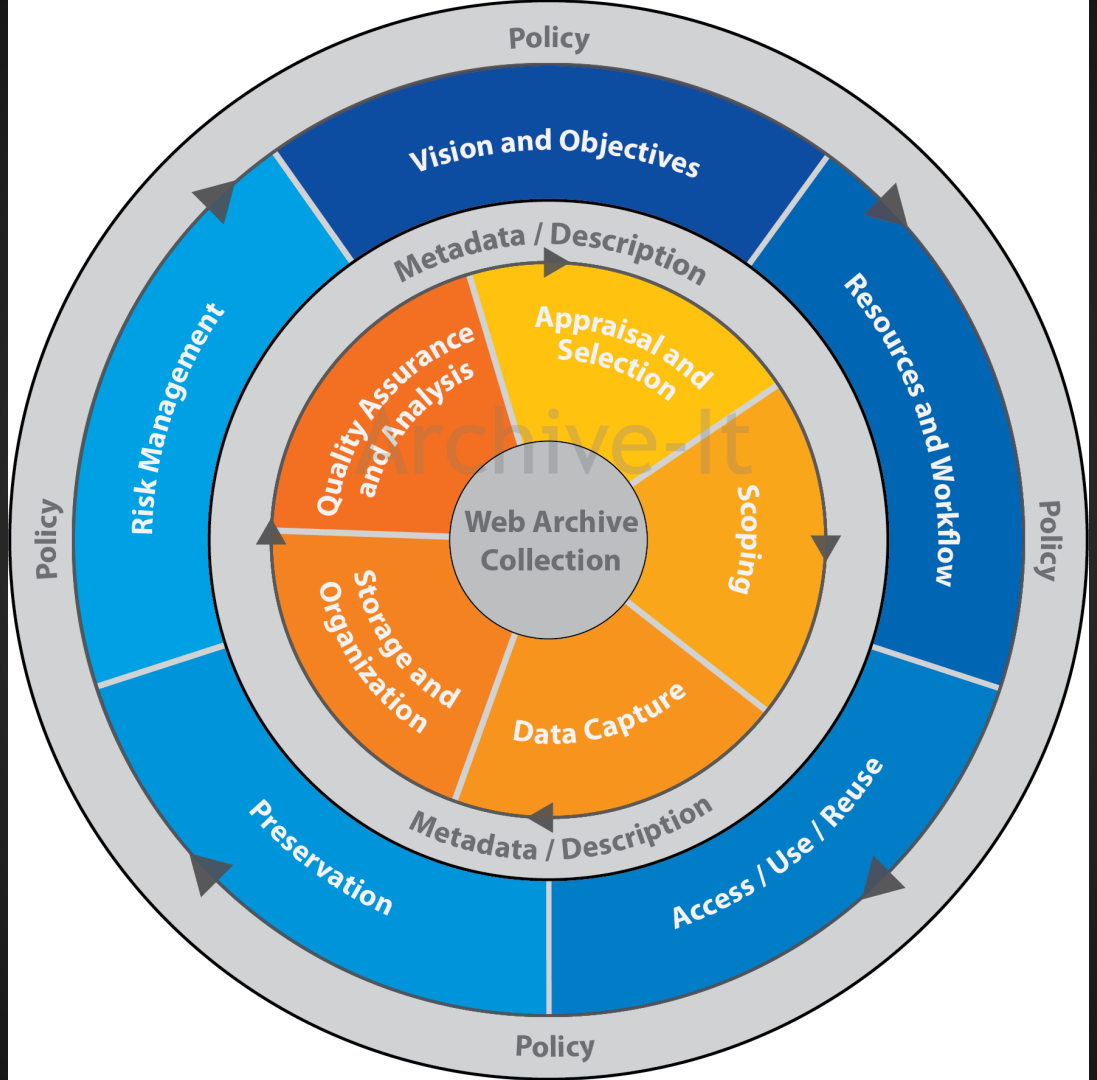


Helps to piece the story together from *massive* web archives

Internet Archive isn’t the only way!

…but they created the Web Archiving Lifecycle


CONCEPTUALISE
CREATE OR RECEIVE
APPRAISE & SELECT
INGEST
PRESERVATION ACTION
STORE
ACCESS, USE & REUSE
TRANSFORM
DISPOSE
MIGRATE
REAPPRAISE
CURATE
PRESERVE
and
and
PRESERVATION PLANNING
DESCRIPTION
REPRESENTATION INFORMATION
COMMUNITY WATCH & PARTICIPATION
Data
(Digital Objects
or
Databases)


Policy
Vision and Objectives
Resources and Workflow
Policy
Access / Use / Reuse
Policy
Preservation
Policy
Risk Management
Policy
Metadata / Description
Appraisal and Selection
Scoping
Data Capture
Metadata / Description
Storage and Organization
Quality Assurance and Analysis
Web Archive Collection
archive-it

Thanks!

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