Getting sh*t done in the digital archives

Anna St. Onge York University
Nick Ruest, York University
MANDATE

We care about stuff like memory/transparency/access/preservation
Why should we care about web content?
The archivist, then, has a unique opportunity in the future if he addresses himself to this vital task of information retrieval, in both the field of historical and modern records. His resources will span the present and the past, and he will hold the key to decision-making and research alike. His bank of interest will be far wider than that of the records manager, but he must learn the language of the computer like his native tongue if he is not to be relegated to the fringe of administration from which he came. We must not be seduced by a kind of academic dolce vita or we will surely die as archivists and will fail to ensure for the future the continuation of that record which has, by so much effort, been saved from the past.
“Our job is to serve the government and the ministers that we report to…

Our job is to keep the institution relevant and to keep it focused on a path where it can still continue to deliver its mandate [with] fewer resources.

The minute we become activists we’re going to lose our judgment and our perspective on how to do that.”

Cecelia Muir, COO of LAC, responding to question about lack of advocacy by LAC leadership regarding cuts in 2012.
"For archivists to abstain from cultural awareness and criticism is tantamount to professional irresponsibility."

Brien Brothman
Anonymous photograph given to author by Melissa Terras via Twitter, 10 June 2012. See: https://twitter.com/melissaterras/status/222703470245978114/photo/1
FOCUS

What should we care about?
current news

Third TEDxYorkU enjoys record attendance
Schulich MBA team wins global real estate case competition
Marathi Dalit writer, intellectual and activist to speak at York
York celebrates innovation in teaching and learning – May 9

University Bookstore makes changes in time for May rush

FEATURED EVENTS

Mothers Speak Up! On Parenting Children with Disabilities - May 6, 9:30-1:30pm

And Social Justice for All: A Student Call for Action - May 8, 9:30-2:30pm

Muscle Health Awareness Day - May 17, 9-4pm

Speakers Series: Let’s Talk - Difficult Conversations (SSDCON) - Jun 21, 8:30-10:30am

more events

YORK IN THE MEDIA

Most people don’t have viable financial plan: poll

Canada’s universal health-care system makes us live democratically

Canada Post issues stamps celebrating Chinatown gates, two GTA locations included

Laptops impede learning as students surf, York U study says
DEMONSTRATE

Showing people we care about stuff and that we can care about their stuff too
GET ‘ER DONE

This is what we did
YUDL Web archiving
https://github.com/yorkulibraries/YUDL-Web-archiving

Islandora Solution Pack Web ARCHive
https://github.com/ruebot/islandora_solution_pack_web_archive
Unhappy Motorist Learns That Her Car Has Failed the Test at an Auto Emission Inspection Station at Norwood, Ohio...08/1975.
Available from: The US National Archives
Public Works Employee at an Auto Emission Inspection Station in Norwood, Ohio...08/1975. Available from: The US National Archives
$ wget --warc-file=foo http://yfile.news.yorku.ca/
Opening WARC file ‘foo.warc.gz’.

Resolving yfile.news.yorku.ca (yfile.news.yorku.ca)... 130.63.173.84
Connecting to yfile.news.yorku.ca (yfile.news.yorku.ca)|130.63.173.84|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: ‘index.html’

OK ............ ................ ............ ................ . 180K=0.2s

2013-05-08 10:17:49 (180 KB/s) - ‘index.html’ saved [32084]
Stop Sign Where Motorists Halt Their Vehicles and Turn Them over to Employees at an Auto Emission Inspection Station in Downtown Cincinnati, Ohio...09/1975.
Available from: The US National Archives.
HTTP/1.1 200 OK
Date: Thu, 05 Sep 2013 06:30:17 GMT
Server: Apache/2.2.14 (Ubuntu)
X-Powered-By: PHP/5.3.2-1ubuntu4.18
Set-Cookie: wfvvt_3604924234=522824fa34b81; expires=Thu, 05-Sep-2013 07:00:18 GMT; path=/
Set-Cookie: spo_171_fa=0125371a1f04cde1ecfd0a10871f4405; expires=Thu, 05-Sep-2013 07:00:18 GMT; path=/
X-Pingback: http://yfile.news.yorku.ca/xmlrpc.php
Keep-Alive: timeout=15, max=100
Connection: Keep-Alive
Transfer-Encoding: chunked
Content-Type: text/html; charset=UTF-8

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<meta name="generator" content="WordPress 3.6" />
<meta name="robots" content="all" />
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-includes/js/jquery/jquery.js?v=1.10.2'></script>
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-includes/js/jquery/jquery-migrate.min.js?v=1.2.1'></script>
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-content/plugins/google-analyticsicator/external-tracking.min.js?v=6.4.5'></script>
<link rel="EditURI" type="application/rsd+xml" title="RSD" href="http://yfile.news.yorku.ca/xmlrpc.php?rsd" />
<link rel="wlwmanifest" type="application/wlwmanifest+xml" href="http://yfile.news.yorku.ca/wp-includes/wlwmanifest.xml" />

<!-- Google Analytics Tracking by Google Analyticator 6.4.5: http://www.videousermanuals.com/google-analyticator/ -->
<script type="text/javascript">
    var analyticsFileTypes = [''];
    var analyticsEventTracking = 'enabled';
</script>
<script type="text/javascript">
    var gaq = gaq || [];
    gaq.push(['_setAccount', 'UA-5080560-1']);
    gaq.push(['_addDevId', 'i9k95']); // Google Analyticator App ID with Google
</script>
Web Archiving Life Cycle
#!/bin/bash
# Bash script that archives selected York University web sites using the WARC ISO format.
PATH=/usr/local/bin:/usr/local/bin:/sbin:/bin:/usr/sbin:/usr/bin
DATE=`date +"%Y-%m-%d"
ISODATE=`date +"%Y-%m-%d"

cd /mnt/DIY/DIY/web-archiving/warcs/yf
mkdir YFILE_$DATE
cd YFILE_$DATE

FILESIZE=`ls -sh YFILE_$DATE | awk '{print $1}'`

xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltopdf --use-xserver --dpi 200 --page-size Letter http://yfile.news.yorku.ca/ YFILE_$DATE.pdf
xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltoimage --use-xserver http://yfile.news.yorku.ca/temp.png
pngcrush temp.png YFILE_$DATE.png
rm temp.png
/usr/local/bin/wget --mirror --page-requisites --warc-file=YFILE_$DATE --no-warc-compression --wait=1 http://yfile.news.yorku.ca/
echo $MODS > YFILE_$DATE.xml
cp YFILE_$DATE.warc /mnt/big_data/web-archiving/wayback-warcs
rm -rf yfile.news.yorku.ca
echo "$(date) - YFile archived" >> /var/log/webarchiving.log
Available from: gifw0rld
Now What!?
Web ARCHive Solution Pack
Description
Daily crawl of YFile

Download
- Warc: YFile - 2013-03-26.warc
- Screenshot: YFile - 2013-03-26.png
- CSV: YFile - 2013-03-26.csv

In collections
- YFile

Details

<table>
<thead>
<tr>
<th>Title</th>
<th>YFile - 2013-03-26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>newsletter--York University</td>
</tr>
<tr>
<td>Description</td>
<td>Daily crawl of YFile</td>
</tr>
<tr>
<td>Publisher</td>
<td>York University</td>
</tr>
<tr>
<td>Contributor</td>
<td>York University's Communications &amp; Public Affairs (Creator)</td>
</tr>
<tr>
<td>Date</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Website</td>
</tr>
<tr>
<td>Format</td>
<td>1.5M, electronic</td>
</tr>
<tr>
<td>Identifier</td>
<td>yul:73040, YFILE_2013_03_26</td>
</tr>
<tr>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>eng</td>
</tr>
<tr>
<td>Relation</td>
<td></td>
</tr>
<tr>
<td>Coverage</td>
<td></td>
</tr>
<tr>
<td>Rights</td>
<td></td>
</tr>
</tbody>
</table>
## Search results

### Query

- "John Greyson and Tarek Loubani"

(1 - 4 of 4)

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Genre</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>YFile - 2013-08-30</td>
<td>Daily crawl of YFile</td>
<td>Website</td>
<td>electronic</td>
</tr>
<tr>
<td>YFile - 2013-08-31</td>
<td>Daily crawl of YFile</td>
<td>Website</td>
<td>electronic</td>
</tr>
<tr>
<td>YFile - 2013-08-28</td>
<td>Daily crawl of YFile</td>
<td>Website</td>
<td>electronic</td>
</tr>
</tbody>
</table>
If you’re going to have to do something, you might as well work with something that interests and motivates you and serves a purpose.
DO BY CARING
Zombie portrait of John Dupuis, by unknown artist. Available from: @dupuisj.
arxivdaleascii

https://github.com/ruebot/arxivdaleascii
#!/bin/bash
# Bash script that archives selected Free Dale Askey web sites using the WARC ISO format.
HOME=/home/nruest/Dropbox/mellen-askey/arxiv/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 -- -s "-screen 0 1280x1024x24" /usr/bin/wkhtmltopdf --dpi 200 --page-size Letter --custom-header 'User-Agent' $line
    /usr/bin/xvfb-run -- -s "-screen 0 1280x1024x24" /usr/local/bin/wkhtmltoimage --use-xserver --custom-header 'User-Agent' $line
    rm tmp.png
    /usr/local/bin/wget --adjust-extension --page-requisites --convert-links --no-parent --random-wait --warc-file=$pad-$DATE $HOME/FDA_$DATE
    zip -r $DATE-$pad.zip $DATE-$pad
    rm -rf $DATE-$pad
    echo "$(date) - $line archived" >> /var/log/daleascii.log
done
http://pinboard.in/u/jbfink/t/mcmaster-temp
http://www.timeshighereducation.co.uk/story.asp?storycode=154569
http://www.safs.ca/academicfreedom/miles.html
http://leiterreports.typepad.com/blog/2013/02/academic-librarian-sued-for-dissing-publisher-in-a-blog-post.html
http://leiterreports.typepad.com/blog/2013/02/shocking-attack-on-academic-freedom-at-mcmaster-by-edwin-mellen-press.html
http://capallibarians.org/2013/02/07/unprecedented-3-million-dollar-lawsuit-filed-against-mcmaster-librarian-for-bloginpost/
http://plgta.org/archives/149
http://sametrosow.wordpress.com/2013/02/08/libel-action-against-mcmaster-librarian-raises-academic-freedom-issues/
http://blogs.princeton.edu/librarian/2013/02/02/a-mcmaster-statement-and-a-petition/
http://gawker.com/5902793/publisher-sues-college-librarian-for-saying-publisher-sucks
http://www.insidehighered.com/news/2013/02/08/academic-librarian-raising-issues-academic-freedom
http://dailynews.mcmaster.ca/worth-mentioning/mcmasters-commitment-to-academic-freedom/
http://www.librarian.net/stax/4033/should-it-be-okay-to-sue-librarians-for-saying-your-books-are-bad/
http://www.philnel.com/2013/02/08/freedaleasky/
http://www.txwaston.com/2013/02/08/textbook-publisher-sues-librarian-for-pointing-out-that-their-books-suck/
https://yuvalib.wordpress.com/2013/02/08/dale-askey/
http://chrisbourg.wordpress.com/2013/02/09/i-hate-rats/
http://africanstudieslibrary.wordpress.com/2013/02/08/mellen-press-sues-librarian-for-libel/
http://boingboing.net/2013/02/08/publisher-launches-3000000.html
http://whymeappleby.wordpress.com/2013/02/09/how-to-respond-to-criticism/
http://acrolog.org/2013/02/09/scholarly-publishing-still-not-making-sense/
http://publishingarchaeology.blogspot.ca/2013/02/watch-out-for-edwin-mellen-press.html
http://www.reddit.com/r/books/comments/185du6/edwin_mellen_press_is_suing_an_academic_librarian/
http://www.reddit.com/r/education/comments/185kn5/an_academic_press_sues_a_librarian_raising_issues/
http://www.ccla.org/rightswatch/2013/02/10/mcmaster-university-accused-of-libel/
http://gavialib.com/2013/02/02/the-librarian-and-the-bully/
"For archivists to abstain from cultural awareness and criticism is tantamount to professional irresponsibility."

"The warning is that if archivists ignore the cultural mandate of their profession, in favor of an administrative point of view and role, the profession will lose vitality."
We are Toronto-area library workers who are concerned with social justice and equality issues, charged with the stewardship of knowledge, championing open access to information, and preserving common space...
The #freedaleaskey Collection

http://freedaleaskey.plggta.org/
Letters of support

ARL and CARL Urge Dropping of Remaining Lawsuit against Askey

ARL-CARL Joint Statement in Support of Dale Askey and McMaster University

Academic Freedom for Librarians: What is it, and why does it matter?

Association of Canadian University Presses Open Letter to the Scholarly Community

Brock University Faculty Association Professional Librarians

George Washington University Libraries Statement

Librarian members of the University of Guelph Faculty Association

McMaster University Faculty Association
Websites

AASUA: Council Passes Motion Supporting Dale Askey

AAUP: AAUP Issues Statement on Edwin Mellen Lawsuit


ACRLog: Scholarly Publishing: Still Not Making Sense

ALA: ALA President Maureen Sullivan responds to Edwin Mellen Press lawsuit

ALPS Yoder: BCLA Press Release on the Dale Askey / McMaster / Edwin Mellen Lawsuit

APBU: APBU Statement of Support for McMaster Associate University Librarian Dale Askey

APLUA: In support of Dale Askey, librarian at McMaster Library – February 22, 2012

ARL: ARL and CARL Urge Dropping of Remaining Lawsuit against Askey

African Studies Library: Mellen Press Sues Librarian for Libel

Agnostic, Maybe: Blind Mellendrama

American Libraries: ALA President Maureen Sullivan responds to Edwin Mellen Press lawsuit
Describing the present
PLG GTA Collections

http://archives.plggta.org/
Collection Collection01 - #FreeDaleAskey collection

Identity area

Reference code | Collection01
Title | #FreeDaleAskey collection
Date(s) | 2012-2013 (Accumulation)
| 1994, 2010-2013 (Creation)
Level of description | Collection

Context area

Name of creator | Progressive Librarians Guild - Greater Toronto Area Chapter (PLG GTA) (2011-
Administrative history | We are Toronto-area library workers who are concerned with social justice and equality issues, charged with the stewardship of knowledge, championing open access to information, and preserving common space. We are interested in issues of freedom of ...

Content and structure area

Scope and content | Collection consists of public statements, publicly available press releases, blog posts, letters of support, online comments and more ephemeral elements related to the #FreeDaleAskey campaign that developed as a response to court proceedings filed by ...

Appraisal, destruction and scheduling | This collection was identified, selected and preserved by members of the Progressive Librarians Guild of the Greater Toronto Area (PLG-GTA) beginning in February 2013. Our selection criteria is based on our aims to identify and preserve online debate ...

Accruals | Websites were captured daily once the presence of a relevant blog post or media story was known by a member of the PLG-GTA.

System of arrangement | Websites have been organized roughly into categories based on the format or genre of the website (i.e. public statements, letters of concern, blog posts, online articles etc.) and entered into the collection collection based on the date site was ...

Conditions of access and use area

Conditions governing access | Materials captured and gathered from various online sources are governed by the access, copyright and reproduction statements attributed by their authors.
Challenges
(pains d’arse)
Contact

Anna St. Onge

Twitter: @deantiquate
Email: astonge@yorku.ca

Nick Ruest

Twitter: @ruebot
Github: ruebot
Email: ruestn@yorku.ca