

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?

NOT REPRODUCED BY THE USE OF OTHER GREAT SOURCES.

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.

Hugh A. Taylor, "Information Retrieval and the Training of the Archivist". *The Canadian Archivist*. 2:3 (1972): 33.

“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."

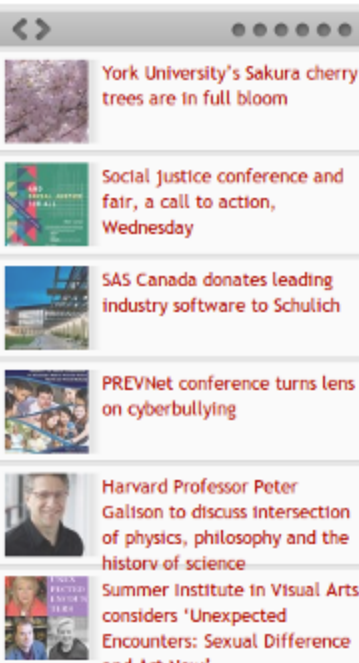
Brien Brothman



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?



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

current news



University Bookstore makes changes in time for May rush

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

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](#)



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.

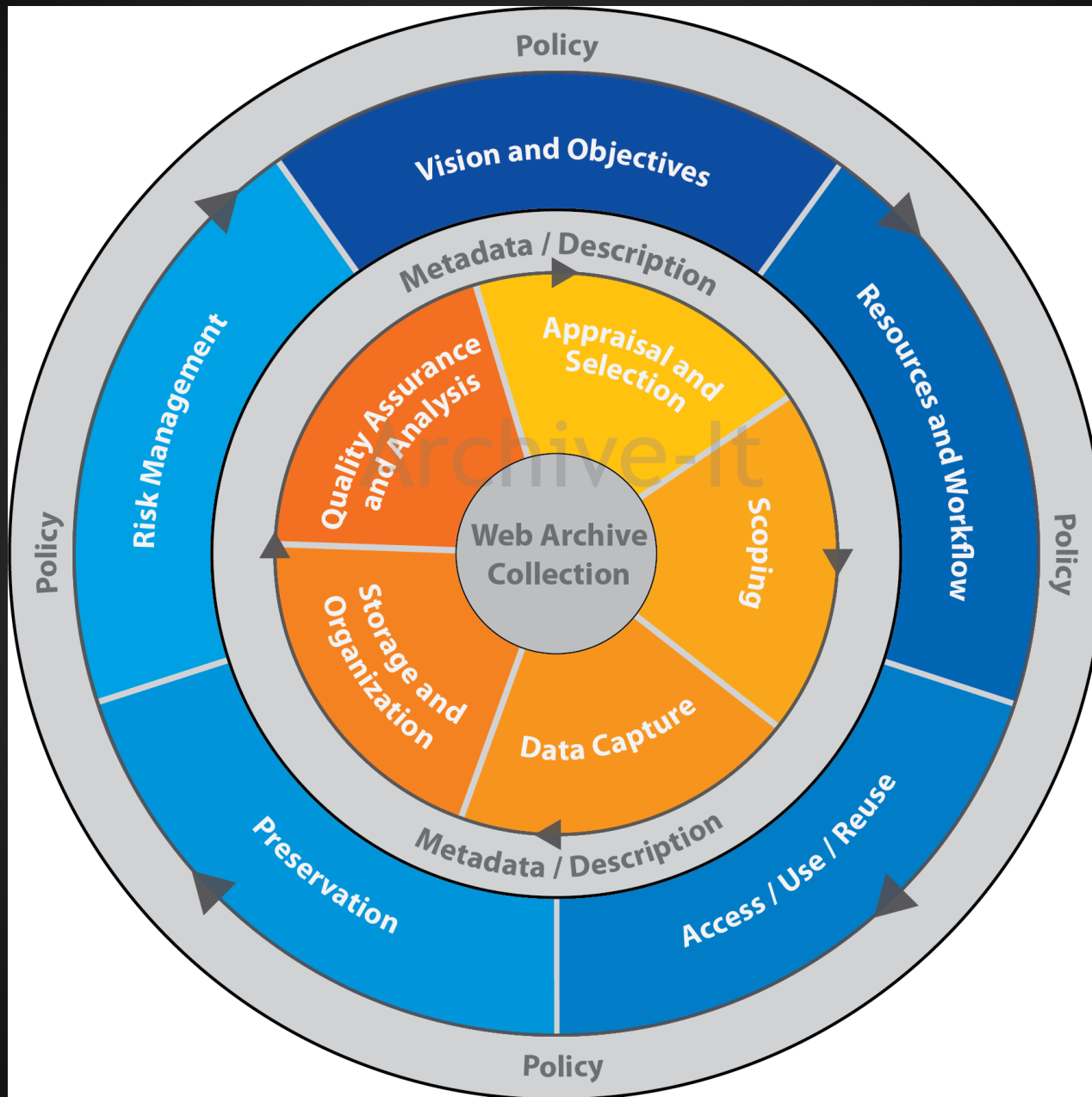
Available from: [The US National Archives](#).

```
WARC/1.0^M
WARC-Type: warcinfo^M
Content-Type: application/warc-fields^M
WARC-Date: 2013-09-05T06:30:17Z^M
WARC-Record-ID: <urn:uuid:ba68aa8d-67f6-4e68-846f-3e6528460552>^M
WARC-Filename: YFILE_2013_09_05.warc^M
WARC-Block-Digest: sha1:W63KEPCYOT7T37FJXU7VKSC3I52QQYS4^M
Content-Length: 306^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=YFILE_2013_09_05" "--no-warc-compres
sion" "--wait=1" "http://yfile.news.yorku.ca/" ^M
^M
^M
^M
WARC/1.0^M
WARC-Type: request^M
WARC-Target-URI: http://yfile.news.yorku.ca/^M
Content-Type: application/http;msgtype=request^M
WARC-Date: 2013-09-05T06:30:17Z^M
WARC-Record-ID: <urn:uuid:7f2b4e94-517f-4a69-9983-26801dfbccd4>^M
WARC-IP-Address: 130.63.173.84^M
WARC-Warcinfo-ID: <urn:uuid:ba68aa8d-67f6-4e68-846f-3e6528460552>^M
WARC-Block-Digest: sha1:WH460WJ65GYHTCGPLKTDSD6RM2XB5UIM^M
Content-Length: 117^M
```

```
HTTP/1.1 200 OK^M
Date: Thu, 05 Sep 2013 06:30:17 GMT^M
Server: Apache/2.2.14 (Ubuntu)^M
X-Powered-By: PHP/5.3.2-1ubuntu4.18^M
Set-Cookie: wfv_t_3604924234=522824fa34b81; expires=Thu, 05-Sep-2013 07:00:18 GMT; path=/^M
Set-Cookie: spo_171_fa=0125371a1f04cde1ecfd0a10871f4405; expires=Thu, 05-Sep-2013 07:00:18 GMT; path=/^M
X-Pingback: http://yfile.news.yorku.ca/xmlrpc.php^M
Keep-Alive: timeout=15, max=100^M
Connection: Keep-Alive^M
Transfer-Encoding: chunked^M
Content-Type: text/html; charset=UTF-8^M
^M
1f8f^M
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">^M
<html xmlns="http://www.w3.org/1999/xhtml">^M
<head>^M
<title>^M
YFile</title> ^M
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />^M
<meta name="generator" content="WordPress 3.6" />^M
<meta name="robots" content="all" />^M
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-includes/js/jquery/jquery.js?ver=1.10.2'></script>
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-includes/js/jquery/jquery-migrate.min.js?ver=1.2.1'></script>
<script type='text/javascript' src='http://yfile.news.yorku.ca/wp-content/plugins/google-analytics/external-tracking.min.js?ver=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
```

Web Archiving Life Cycle



```
#!/bin/bash
# Bash script that archives selected York University web sites using the WARC ISO format.
PATH=/usr/local/sbin:/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}'`
MODS="<mods:titleInfo><mods:title>YFile - $ISODATE</mods:title></mods:titleInfo><mods:name type='corporate'><mods:namePart>York University's Communications & Public Affairs</mods:namePart><mods:role><mods:roleTerm authority='marcrelator' type='text'>Creator</mods:roleTerm></mods:role></mods:name><mods:typeOfResource>mixed material</mods:typeOfResource><mods:genre authority='lctgm'>Website</mods:genre><mods:originInfo><mods:dateIssued>$ISODATE</mods:dateIssued><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><mods:identifier type='local'>YFILE_$DATE</mods:identifier><mods:physicalDescription><mods:form authority='marcform'>electronic</mods:form><mods:extent>$FILESIZE</mods:extent></mods:physicalDescription><mods:subject><mods:topic>newsletter</mods:topic><mods:topic>York University</mods:topic></mods:subject></mods:mods>"

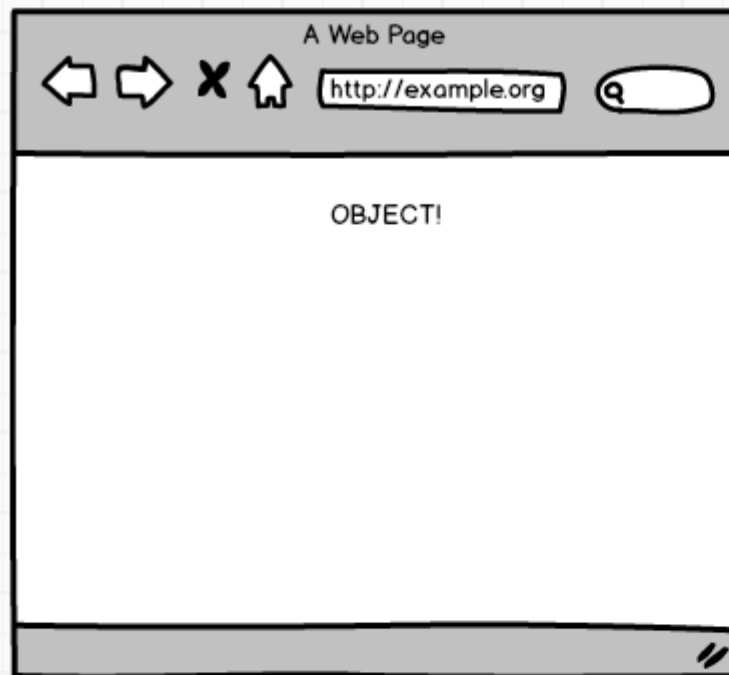
xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltopdf --use-xserver --dpi 200 --page-size Letter
http://yfile.news.yorku.ca/ YFILE_$DATE.pdf
xvfb-run -a -s "-screen 0 1280x1024x24" wkhtmltoimage --use-xserver http://yfile.news.yorku.ca/
temp.png
pngcrush temp.png YFILE_$DATE.png
rm temp.png
/usr/local/bin/wget --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
```



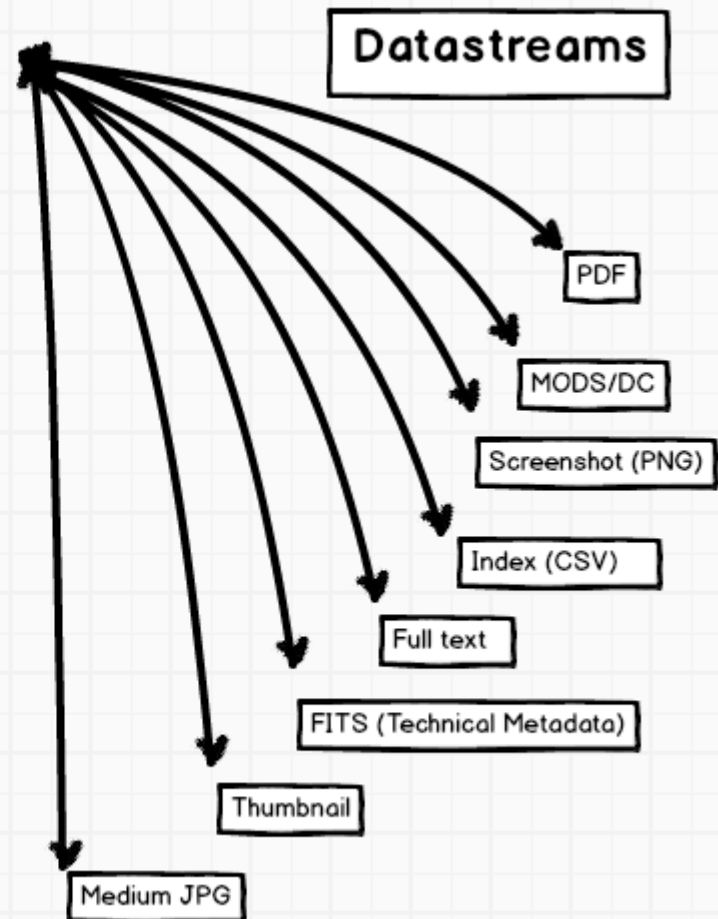
Available from: [gifw0rld](https://www.gifworld.com)

Now What!?

Web ARChive Solution Pack



**Archival Object -
WARC (ISO 28500)**



Websites

1 2 3 4 5 next



[AAUP: AAUP Issues Statement on Edwin Mellen Lawsuit](#)



[ACLRinsider: ACRL Board of Directors Statement on Edwin Mellen Press v. Dale Askey](#)



[ACRLog: Scholarly Publishing: Still Not Making Sense](#)



[ALA: ALA President Maureen Sullivan responds to Edwin Mellen Press lawsuit](#)



[ALPS Yodeler: 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 Melliendrama](#)



[American Libraries: ALA President Maureen Sullivan responds to Edwin Mellen Press lawsuit](#)



[Andrew Petter, SFU President and Vice-Chancellor: In defence of academic freedom](#)





YFile - 2013-03-04



YFile - 2013-03-05



YFile - 2013-03-06



YFile - 2013-03-07



YFile - 2013-03-08



YFile - 2013-03-09



YFile - 2013-03-10



YFile - 2013-03-11



YFile - 2013-03-12



YFile - 2013-03-13



YFile - 2013-03-14



YFile - 2013-03-15



[illegible]

Description

Daily crawl of YFile

Download

- Warc: [YFile - 2013-03-26.warc](#)
- PDF: [YFile - 2013-03-26.pdf](#)
- Screenshot: [YFile - 2013-03-26.png](#)
- CSV: [YFile - 2013-03-26.csv](#)

In collections

- [YFile](#)

Details

Title	YFile - 2013-03-26
Creator	
Subject	newsletter--York University
Description	Daily crawl of YFile
Publisher	York University
Contributor	York University's Communications & Public Affairs (Creator)
Date	
Type	Website
Format	1.5M, electronic
Identifier	yul:73040, YFILE_2013_03_26
Source	
Language	eng
Relation	
Coverage	
Rights	

Search results

Query

- (-) "John Greyson and Tarek Loubani"

(1 - 4 of 4)



Title	YFile - 2013-08-30
Description	Daily crawl of YFile
Genre	Website
Type	electronic



Title	YFile - 2013-08-31
Description	Daily crawl of YFile
Genre	Website
Type	electronic



Title	YFile - 2013-08-28
Description	Daily crawl of YFile
Genre	Website
Type	electronic

search

Subject

- [York University](#) (4) + -
- [newsletter](#) (4) + -

Genre

- [Website](#) (4) + -

Publisher

- [York University](#) (4) + -

Language

- [English](#) (4) + -

Format

- [mixed material](#) (4) + -

LEARN BY DOING

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




Herbert Richardson in his office at Edwin Mellen Press in Lewiston, N.Y.
Photograph by Jason Greene, 2013.
Available from: [The Chronicle of Higher Education](#)



Zombie portrait of John Dupuis, by unknown artist. Available from:
[@dupuisj](https://twitter.com/dupuisj).

arxivdaleascii

<https://github.com/ruebot/arxivdaleascii>

 ruebot 3 months ago so, uhhh, nazis, eh?

2 contributors  



executable file | 27 lines (24 sloc) | 1.304 kb

Edit

Raw

Blame

History

Delete

```

1  #!/bin/bash
2  # Bash script that archives selected Free Dale Askey web sites using the WARC ISO format.
3  HOME=/home/nruest/Dropbox/mellen-askey/arxiv/sites
4  DATE=`date +"%Y_%m_%d"`
5  SITES=/home/nruest/git/arxivdaleascii/arxivdaleascii-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-A
18     /usr/bin/xvfb-run -a -s "-screen 0 1280x1024x24" /usr/local/bin/wkhtmltoimage --use-xserver --custom-header 'User-Agent' '
19     /usr/bin/pngcrush tmp.png $pad-$DATE.png
20     rm tmp.png
21     /usr/local/bin/wget --adjust-extension --page-requisites --convert-links --no-parent --random-wait --warc-file=$pad-$DATE
22     cd $HOME/FDA_$DATE
23     zip -r $DATE-$pad.zip $DATE-$pad
24     rm -rf $DATE-$pad
25     echo "$(date) - $line archived" >> /var/log/daleascii.log
26 done

```



ruebot 7 days ago fixes ruebot/arxivdaleascii#2

1 contributor

file | 266 lines (265 sloc) | 21.0 kb

[Edit](#) [Raw](#) [Blame](#) [History](#) [Delete](#)

```
1 http://web.archive.org/web/20110630153231/http://htwkbk.wordpress.com/2010/09/22/the-curious-case-of-edwin-mellen-press/
2 https://www.change.org/petitions/edwin-mellen-press-end-libel-suit-against-dale-askey-and-mcmaster-university
3 http://leiterreports.typepad.com/blog/2013/02/best-philosophy-publishers-in-english.html
4 https://pinboard.in/u:jbfinck/t:mcmaster+emp
5 http://chronicle.com/forums/index.php/topic,122855.0.html
6 http://www.timeshighereducation.co.uk/story.asp?storycode=154569
7 http://www.safs.ca/academicfreedom/miles.html
8 http://blogs.princeton.edu/librarian/2013/02/edwin-mellen-press-suing-a-librarian
9 http://lawprofessors.typepad.com/law_librarian_blog/2013/02/academic-librarian-sued-for-dissing-publisher-in-a-blog-post.ht
10 http://leiterreports.typepad.com/blog/2013/02/shocking-attack-on-academic-freedom-at-mcmaster-by-edwin-mellen-press.html
11 http://capalibrarians.org/2013/02/07/unprecedented-3-million-dollar-lawsuit-filed-against-mcmaster-librarian-for-blogpost/
12 http://plggta.org/archives/149
13 http://samtrosow.wordpress.com/2013/02/08/libel-action-against-mcmaster-librarian-raises-academic-freedom-issues/
14 http://blogs.princeton.edu/librarian/2013/02/a-mcmaster-statement-and-a-petition/
15 http://leiterreports.typepad.com/blog/2013/02/a-statement-by-mcmaster-on-the-mellen-press-lawsuit.html
16 http://gawker.com/5982793/publisher-sues-college-librarian-for-saying-publisher-sucks
17 http://www.insidehighered.com/news/2013/02/08/academic-press-sues-librarian-raising-issues-academic-freedom
18 http://www.the-digital-reader.com/2013/02/08/how-not-to-silence-critics-canadian-publisher-sues-librarian-for-defamation/
19 http://dailynews.mcmaster.ca/worth-mentioning/mcmasters-commitment-to-academic-freedom/
20 http://chronicle.com/blogs/wiredcampus/edwin-mellen-press-sues-university-librarian-for-libel/42193
21 http://www.librarian.net/stax/4033/should-it-be-okay-to-sue-librarians-for-saying-your-books-are-bad/
22 http://www.philnel.com/2013/02/08/freedaleaskey/
23 http://www.txwatson.com/2013/02/08/textbook-publisher-sues-librarian-for-pointing-out-that-their-books-suck/
24 http://www.eduhacker.net/libraries/edwin-mellen-press-sues-librarian.html
25 https://yufalib.wordpress.com/2013/02/08/dale-askey/
26 http://chrisbourg.wordpress.com/2013/02/09/i-hate-rats/
27 http://africanstudieslibrary.wordpress.com/2013/02/08/mellen-press-sues-librarian-for-libel/
28 http://boingboing.net/2013/02/08/publisher-launches-3000000.html
29 http://distlib.blogs.com/distlib/2013/02/edwin-mellen-press-files-lawsuit-against-academic-librarian.html
30 http://jwhyteappleby.wordpress.com/2013/02/09/how-to-respond-to-criticism/
31 http://www.metafilter.com/124794/On-questioning-the-quality-of-a-publisher
32 http://acrlog.org/2013/02/09/scholarly-publishing-still-not-making-sense/
33 http://publishingarchaeology.blogspot.ca/2013/02/watch-out-for-edwin-mellen-press.html
34 http://www.slaw.ca/2013/02/09/edwin-mellen-presss-curious-case/
35 http://www.reddit.com/r/books/comments/185do6/edwin_mellen_press_is_suing_an_academic_librarian/
36 http://www.reddit.com/r/education/comments/185kns/an_academic_press_sues_a_librarian_raising_issues/
37 http://www.ccla.org/rightswatch/2013/02/10/mcmaster-university-accused-of-libel/
38 http://www.insidehighered.com/blogs/world-view/speak-out-and-get-sued
39 http://gavialib.com/2013/02/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



Brock University Library Support the Academic Freedom



QUFA Statement



Librarian members of the University of Guelph Faculty Association



McMaster University Faculty Association

Websites

1 2 3 4 5 next



[AASUA: Council Passes Motion Supporting Dale Askey](#)



[AAUP: AAUP Issues Statement on Edwin Mellen Lawsuit](#)



[ACLRinsider: ACRL Board of Directors Statement on Edwin Mellen Press v. Dale Askey](#)



[ACRLog: Scholarly Publishing: Still Not Making Sense](#)



[ALA: ALA President Maureen Sullivan responds to Edwin Mellen Press lawsuit](#)



[ALPS Yodeler: 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/>



Creator(s)

- Progressive Librarians Guild - Greater Toronto Area Chapter (PLG GTA)

Digital object



Results 1 to 10 of 220

[See all](#)

Collection

▼ Collection Collection01 - #FreeDaleAskey ...

- Series S01 - Letters of support
- Series S02 - Blog posts in response to co...
- Series S03 - Twitter feeds (?)
- Series S04 - Original blog posts (?)
- Series S05 - Related blog posts and articles

...

Import

- XML
- CSV

Export

- Dublin Core 1.1 XML
- EAD 2002 XML

Browse

- Archival descriptions
- Authority records
- Archival institutions
- Functions
- Subjects
- Places

[View archival description](#)

Collection Collection01 - #FreeDaleAskey collection

Identity area

Reference code	Collection01
Title	#FreeDaleAskey collection
Date(s)	<ul style="list-style-type: none">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