

Digital Studies / Le champ numérique, 2016

The great WARC adventure: Using SIPS, AIPS, and DIPS to document SLAPPs

Alphabetical author list:

Ian Milligan, University of Waterloo: i2milligan@uwaterloo.ca

Nick Ruest, York University: ruestn@yorku.ca

Anna St.Onge, York University: astonge@yorku.ca

Peer-reviewed by: Léon Robichaud, Université de Sherbrooke; Frédéric Clavert, University of Lausanne

Abstract / Résumé

This paper outlines the circumstances surrounding a libel case that was filed against academic librarian Dale Askey by publisher Herbert Richardson and his company Edwin Mellen Press, the resulting online debate, protest, and advocacy, and the effort by a small team to capture, preserve, and make available preserved websites related to the event. Ruest, a programmer and archivist-librarian, presents the technical aspects of capturing and preserving WARC files. St.Onge, an archivist, reflects on some of the challenges of creating a traditional finding aid to contextualize and provide access to the collected electronic content. Milligan, a historian, discusses some preliminary findings based on analysis of the data set. Finally, the authors reflect on the issues brought to the surface by their engagement with questions of academic freedom, librarianship, and public advocacy and on how smaller groups of like-minded professionals can preserve online material whose afterlife might otherwise prove fleeting.

Cet article expose les circonstances entourant une affaire de diffamation déposée contre le bibliothécaire universitaire Dale Askey par l'éditeur Herbert Richardson et sa compagnie Edwin Mellen Press, les débats, protestations et plaidoyers en ligne qui en ont résulté, et les efforts d'une petite équipe pour tenter de capturer et de préserver les sites Web reliés à l'événement et en permettre l'accès. Ruest, un programmeur et archiviste/bibliothécaire, présente les aspects techniques reliés à l'acquisition et la préservation des fichiers d'archivage Web (WARC). St.Onge, un archiviste, examine certains des défis visant à créer un instrument de recherche traditionnel pour mettre en contexte le contenu électronique recueilli et en fournir l'accès. Milligan, un historien, discute de certaines constatations préliminaires fondées sur l'analyse de l'ensemble des données. Enfin, les auteurs analysent les questions soulevées par leur engagement en termes de liberté académique, de bibliothéconomie et de plaidoyer public, et les moyens par lesquels un groupe plus restreint de professionnels, qui partagent les mêmes convictions, peut préserver des documents en ligne dont les chances de survie pourraient autrement s'avérer éphémères.

Keywords / Mots Clés

OAIS, archives, digital repositories, digital history, text analysis, text mining, #freedaleaskey

Contents / Contenu

Introduction

The story of Dale Askey a.k.a. #freedaleaskey

Taking the gamble: Capturing historical moments for posterity

Putting the archive together

Capturing two hundred pages in a single shell script (creating SIPS)

Islandora solution pack Web ARChive: AIP/DIP dissemination

Arrangement

Islandora repository

Storage and organization using Fedora commons & Islandora

Intellectual arrangement and creation of a traditional archival finding aid

Employing OS descriptive software

Reflection on archival practice

A historian in the archive

Articulating a (best) practice

Afterword

Works cited / Liste de références

Introduction

On 11 February 2013, Rick Anderson, a librarian and columnist for the blog *The Scholarly Kitchen*, posted a detailed story of his most recent interaction with the president of the Edwin Mellen Press, a scholarly publisher that had recently accused another librarian, Dale Askey, of libel. The company had filed a civil case against both Askey and his employer, McMaster University in Hamilton, Ontario, suing for a combined \$4.25 million.

The next month, Anderson's initial post and a related post were removed from *The Scholarly Kitchen*. On 29 March 2013, Anderson posted an announcement that the Society for Scholarly Publishing, the site's publisher, had received a letter from a lawyer representing Edwin Mellen Press (EMP) founder Dr. Herbert E. Richardson, demanding that the society remove two offending posts because they contained "disparaging comments about our [EMP's] publishing program, the quality of our books, and attacked the character of our editor, Professor Herbert Richardson" (Amendola 2013; Anderson 2013). The posts were reinstated 3 April 2013 by the board of directors of the SSP, accompanied by a lengthy rationale by SSP president Carol Anne Meyer. It became clear that a struggle for public opinion was being waged between EMP's and Askey's supporters.

In such a rapidly evolving environment, we needed to act to ensure this material could be saved. Though web sources offer readers the advantage of nearly instantaneous updates to a developing story, they are, by their very nature, ephemeral. They can be edited, changed, and removed at any point. In the case of EMP vs. Askey, where the struggle to sway public opinion was played out in a public forum, there was concern that edits to websites, cease and desist demands, and other efforts to control the narrative would—in a very short period of time—remove information from the historical record. As discussions evolved at *Inside Higher Ed* and across multiple web forums of the *Chronicle of Higher Education*, and as takedown notices began to proliferate, the risk of losing material grew. Once lost, the only evidence that would survive would be the hearsay of site administrators and blog authors. We deemed it wise to offer a full accounting of the documentation in order to give the public access to as many sides of the developing story as possible.

The suit against Dale Askey had the potential to be a milestone case in scholarly communication, revolving as it did around central questions of academic freedom, the role of blogging in librarianship, the nature of publishing in the digital age, and the professional practice of librarians. But, given the vulnerability of the historical record in this situation, what would survive as evidence? Although the Internet Archive's extensive scrapes of the World Wide Web might catch some of this content, sites are preserved only every couple of months. The Internet Archive's infrastructure is not set up to capture rapidly changing web pages. Even if the archive could be calibrated to focus on moments of rapid change, the Dale Askey case was not a high-profile event like that of the recent Russian invasion of Crimea, during which the Internet Archive's Archive-It service captured and preserved material from a local media outlet as their offices were being ransacked (see, e.g., Milligan 2014; Dewey 2014; Taylor 2014; Kaplan 2014). Traditionally, web archiving has been dominated by large institutions, from the San Francisco-based Internet Archive to individual universities to national libraries such as Library and Archives Canada. However, in the rapidly evolving case of Dale Askey, whose context was relatively local, intervention by any of these institutions was not feasible. Someone within the community had to take action.

The Progressive Librarians Guild's Greater Toronto Area Chapter (PLG-GTA), of which some of the authors are members, set a plan into action. PLG-GTA is an organization of Toronto-area library workers concerned with social justice and equality, championing open access to information and the preservation of common space (PLG-GTA 2015). The approach involved the deployment of three interconnected programs: one that captured and preserved specific web pages related to the Dale Askey case, one that generated PDFs of the sites, and one that created long screenshots of the page content in a PNG file. Every day, these programs queried an extensive (and growing) list of relevant sites. With a minimum of time and little cost beyond storage, PLG-GTA gathered a large corpus of material relating to this story. This event-based collection, consisting of 226 digital objects, each of which contained daily snapshots of the examined sites over several months, offers an invaluable record of the online debate about the court case. The Islandora Web ARChive (WARC) solution pack (which combines the three programs listed above) provides an ideal research interface, creating WARC objects, PDFs, and PNG derivatives from a very broad capture of content. As Islandora is an open-source digital asset management platform based on Fedora (repository) and Drupal, the WARC solution pack is a Drupal module that integrates well with Islandora, allowing users to deposit, create derivatives, and interact with a given type of object. Ease of use makes it ideal for adoption by organizations with minimal infrastructure. After the creation of the #freedaleaskey collection, PLG-GTA created a finding aid to facilitate access and to document the team's intellectual process.

The first part of this article outlines the collection, acquisition, and archival description of the #freedaleaskey collection, while the second part revolves around how historians might use and interpret it. It is our argument that this event-based web archiving procedure is worth consideration by institutions. From this archive, we can trace how the story shifted emphasis: from one of libel and academic freedom in the first weeks following the announcement of the suit, to a broader discussion about scholarly publishing, and finally to examinations of the history of Edwin Mellen Press founder Herbert Richardson. As we enter an era of digital history—a nebulous subfield of the historical profession that encompasses the use of digital tools both to reach new publics and to do history in new ways—this approach to scholarship can also serve as an example of the possibilities of research in the age of the World Wide Web. By providing a rich data set of daily captures of the identified sites (at a modest cost in terms of infrastructure and labour using open-source tools), both close reading and more macroscopic data mining methodologies are possible.

The story of Dale Askey a.k.a. #freedaleaskey

Dale Askey is an academic librarian. He is currently employed as associate university librarian responsible for library and learning technologies and administrative director of the Lewis & Ruth Sherman Centre for Digital Scholarship, at McMaster University in Hamilton, Ontario. In 2010, however, Askey was a subject librarian working at Kansas State University. As part of his responsibilities there, drawing on his subject specialization expertise as a graduate degree holder in Germanic languages and literatures, Askey supervised purchasing and collection development for arts and humanities. In tandem with these professional responsibilities, Askey wrote and managed an active academic blog, *Eintauchen: dive in*, where he discussed many aspects of his activities as a professional academic librarian.

On 22 September 2010, Askey wrote a blog post discussing the scholarly output of the Edwin Mellen Press (EMP). In it he questioned the quality of the scholarship published by the press, the high prices EMP charged for its titles, and the wisdom of purchasing such titles for academic libraries in a climate of shrinking budgets (Askey 2010a). The public discussion that followed

between Askey and various readers (many of them authors published by EMP) ranged in scope, but throughout his conversations with readers, Askey reiterated his main criticisms:

The titles are nearly always too narrow in scope/too marginal (as you said, journal articles would make sense here), the texts are not professionally edited, the physical quality is suspect, and the prices are too high ([Askey 2010b](#)).

It was not until mid-2012, after Askey had settled into his position at McMaster University, that he received formal notice that EMP's president, Dr. Herbert E. Richardson, was suing him for libel. In December 2012, both Askey and his employer, McMaster University, were served with a second suit naming EMP and Richardson as plaintiffs. In both suits, EMP and Richardson alleged that Askey's post back in 2010, and his refusal to censor or delete public comments to the site, constituted libel under Canadian law.

These suits, which some ([Robinson 2013](#)) have defined as SLAPPs (Strategic Lawsuit Against Public Participation), generated a great deal of online discussion and debate when they became public knowledge in February 2013. Because these and subsequent conversations could factor into any resulting legal proceedings and because of the issues of freedom of speech, academic freedom, and publishing they raised, the PLG-GTA decided to document and preserve them for the historical record. The authors of this paper participated in the project team.

The task of preserving would prove a challenge. As we will discuss, some websites were pressured to take down their blog posts (at least briefly, as we saw above from the actions of *The Scholarly Kitchen*; some websites went dark (as in the case of an essay posted by Murray Miles); and vexingly, the Canadian Association of Professional Academic Librarians changed their URL link structures without providing automatic redirects to their old websites. The Association of Research Libraries (ARL) also migrated their site without setting up redirects, but added them after members of the PLG-GTA asked them to do so. The evolution in online records related to Dale Askey's case, then, presents a problem. How can these records be preserved in their original forms?

Taking the gamble: Capturing historical moments for posterity

The initial impetus for this grassroots project to preserve materials around #freedaleaskey came from within York University, in Toronto, Canada. Nick Ruest was working out a practical solution for web harvesting when the Dale Askey case became public within the wider library community. We saw an opportunity to respond to a pressing public need and test workplace archival practice.

Dale Askey's situation was an ideal case study since much of the public discourse surrounding the case took place on a wide range of platforms: academic blogs, library-related discussion groups, and social media forums (predominantly Twitter). In a constant state of flux, this public discourse even included instances of subterfuge. If the PLG-GTA waited for the issues to settle, important elements and details would be lost, deleted, or edited for posterity. The PLG-GTA team needed to get into the thick of it to document what was happening at the peak of the controversy.

One element that project team members wished to capture was the subtle changes in online conversations that were developing as a result of public interest in the Dale Askey defamation case. Of particular concern were some of the campaigns underway in the comment fields of many blog posts. We wanted to document them before they were edited or deleted. Also, as a profession with a vested interest in access to and free exchange of information, we had a general interest in the story that was breaking online and the conversation that resulted from it.

The PLG-GTA wanted to preserve this information. But how? Currently, institutions dominate web archiving activities. This is unsurprising given the traditional computational, connectivity, and storage needs of web archiving. Web archiving emerged out of a mid-1990s fear around digital preservation: data loss, from changes in software (files becoming truly multimedia), storage costs, copyright challenges, and rapidly evolving media, was tempering the early promise of a record of everything ([Lesk 1995](#); [Kuny 1997](#)). In 1996, the first major web archive initiatives appeared. The non-profit Internet Archive, founded that year by Internet entrepreneur Brewster Kahle ([1996](#)), aimed to conduct comprehensive crawls of the extant World Wide Web; in this it expanded upon much smaller crawls by the Smithsonian Museum and University of North Texas in relation to federal American politics. The same year, the National Library of Australia sought to preserve "selected Australian online publications, such as electronic journals, government publications, and web sites of research or cultural significance" ([National Library of Australia 2009](#)), and Sweden's Royal Library launched Kulturarw3, which began comprehensive crawls of the Swedish Web, completing seven between 1996 and 2000 ([Arvidson, Persson, and Mannerheim 2000](#)). The number of institutions participating in such activities has expanded. The International Internet Preservation Consortium ([2015](#)), chartered in 2003 by twelve members, has now grown to fifty national libraries, service providers, and non-profit foundations. National libraries, including the British Library ([2014](#)) and the Bibliothèque nationale de France ([2006](#)), have legal deposit regulation, allowing them to conduct national-level crawls of their countries' websphere (not just the country-specific top-level domain but also websites registered and based within national boundaries). Scrape frequency is generally once a year for most websites, with more curated subject collections receiving bi-annual, monthly, weekly, or even daily (often just the front page) snapshots. The periodic nature of these scrapes means that sometimes groups need to take web archiving into their own hands if they want to capture at a sufficiently detailed level an event they deem significant.

At this point, it would perhaps be best to acknowledge considerable ambiguity in the use of the term *archive* within the context of our professional and scholarly activities. Other scholars have elegantly articulated the matter (see [Theimer 2012](#); [Owens 2014a](#)). Whereas the content we captured, preserved, and "archived" would most likely be categorized as a collection in the context of an archival institution (i.e. a collection defined as an assemblage of materials from a variety of sources), within the context of online website preservation the term "archive" tends to summarize the act of capture, documentation, and preservation of online content. As well, the ubiquity of "archive" as a verb to describe filing, preserving, or simply not deleting accumulated online content (emails, tweets, RSS feeds) make the clear definition of *our* particular use of the term a moot point. For the most part, a general audience recognizes the word "archive" as a signifier for our general activities related to the project. In the spirit of productive collaboration, the co-authors decided to limit debate on how to use and characterize the term and instead created a crosswalk of unfamiliar terms or terms with multiple meanings that would facilitate our work together.

As the Dale Askey case demonstrates, responsibility for web archiving cannot be left entirely to large state institutions. National

collections constituted through legal deposit and resources like the Internet Archive can provide high-level context but case studies are necessary to help flesh out a more precise picture of life on the Web. With annual, or even monthly, scrapes, archivists may not be able to capture the granular detail that historians require to weave their stories. If a comment is posted and subsequently removed in the interim, it is gone. The same goes for re-edited blog posts and the flow and development of discussions. Virtually none of the online activity and debate occurring around the Dale Askey case would have been captured in a large top-level crawl of major websites; anything captured would have been happenstance. Currently, our national institution, Library Archives Canada, conducts limited web archiving focused on the activities of select federal departments as well as select topic-based crawls of Canadian sites. Grassroots initiatives are needed to complement the formal institutional programs that are more focused on broad sweeps of Canadian domains online.

But what shape should these grassroots initiatives take? One obvious forum is from within an academic library, which enjoys information technology support and forms an integral part of the scholarly mission. Like many other universities, York University (2015) has record retention schedules to facilitate the transfer of official university records of lasting historical value to the archives. Yet a number of factors have made implementing similar practices in the online environment challenging: distributed IT systems, lack of naming conventions, time pressures, office operations silos, and reorganizations that leave websites abandoned or neglected, among others. It is our experience that organizational cultures do not necessarily view websites as official documents, but rather as transactional, fluid spaces. This attitude persists despite the fact that such spaces are increasingly the home of important content that has lasting historical and archival value. So rather than wait for policies to roll out through the usual bureaucratic process, York University Libraries began to preserve a selection of university websites and communications that documented York's campus activities in a systematic and easily accessed manner.

York University Libraries were thus already engaged in preserving their own official websites when the Dale Askey case broke. It gave our team the opportunity to readjust the tools already in use to document a wider social movement that was shaping our collective experience as academic librarians and generating considerable online debate. The project was organized fairly quickly; the aim was to capture sites daily to create a permanent record of the unfolding story and to preserve sites and commentary that were changing over time or being actively deleted. The goal was to document the story of the lawsuit itself and the public reaction to this unprecedented development in the history of academic libraries.

Putting the archive together

The PLG-GTA team had a set of objectives, a range of skill sets, and the urgency of an unfolding story as motivation. But how did things actually get done? The best practices framework for an Open Archival Information System (OAIS) served as the starting point to generate three important information packages that could form the basis of a future data set for study: Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP). For those unfamiliar with OAIS, a SIP is a set of data that a researcher creates and sends to the archive. In the case of this project, the SIP is the daily output of the collection shell script we describe below. An AIP is a SIP, which may or may not be transformed, that is stored by the archive. In the case of this project, the AIP is the transformed object that is stored by the #freedaleaskey archive. The DIP is the object or derivative(s) of the object that is provided to a user when the object is requested. In the case of this project, the DIP is the group of data streams of the object (e.g., an image of the website as one large PNG or PDF or a copy of the WARC file itself) that is provided on a given crawl's page. DIPs are typically in formats that are more compact and more easily manipulated or queried by researchers, whereas the AIP is kept secure and audited to ensure its integrity over time. In order to make the rationale behind the PLG-GTA's decisions transparent, we will outline the following processes: the process and programs developed to create the collection's SIPs; the decision to use Islandora to manage and disseminate AIPs and DIPs; and finally, the decision to highlight "hot zones" of activity, areas of concern, and interconnections between posts through an intellectual arrangement and a traditional archival finding aid.

Capturing two hundred pages in a single shell script (creating SIPs)

The release of Wget 1.14 (a command-line utility for getting files from the Web) included WARC functionality, allowing the PLG-GTA team to capture and preserve over two hundred blog entries every night using twenty-five line shell scripts (for a discussion of Wget, see Milligan 2012). Indeed, our use of capture is deliberate. It is interesting how *Capture*, in the online environment, mirrors a series of more complex tasks that encompass the archival activities of appraisal and acquisition of documents in more traditional contexts. The structure of the script can set the parameters that determine the level of detail when the original state of a particular site is documented, but often this process appears generic (or worse, invisible) to the end user. While automated (i.e., scripted), it is by no means neutral and in fact requires the eye of an archivist who can make informed decisions about what is to be captured and how. Far from being an obsolete skill, the ability of archivists to make selection and appraisal choices is more relevant than ever. What has changed is the environment and tools required to operate effectively in an online environment.

We capture these records in WARC format. WARC functionality has been a boon to the capture and preservation of websites. We can now capture a website simply by running the command

```
wget --warc-file=boycottmellenpress http://boycottmellenpress.org
```

The results look like this:

```
$ wget --warc-file=boycottmellenpress http://boycottmellenpress.com
Opening WARC file 'boycottmellenpress.warc.gz'.
--2014-08-25 14:14:03-- 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% 3.98M=0.001s
2014-08-25 14:14:03 (3.98 MB/s) - 'index.html' saved [3004/3004]
```

This command-line utility can be called in a shell script, allowing a user to run multiple Wget commands without having to manually type each command separately. The PLG-GTA team was therefore able to automate the capture and preservation of a list of web pages. They accomplished this with the script [arxivdaleascii](https://github.com/ruebot/arxivdaleascii/blob/master/arxivdaleascii) (available at <https://github.com/ruebot/arxivdaleascii/blob/master/arxivdaleascii>), which iterated a list of over two hundred pages, now found at <https://github.com/ruebot/arxivdaleascii/blob/master/arxivdaleascii-sites.txt>. When run, the script does a number of things. First, it grabs the current date from the system with the command

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

Second, it creates a directory based on that date and changes to that directory:

```
mkdir FDA_$(DATE)
cd FDA_$(DATE)
```

It then reads a line-delimited text file containing a list of sites (this means each line contains a website address, as does the second line, and so on), and begins iterating over each line with this command:

```
cat $SITES | while read line; do
```

For each line, a few more commands are run, first creating an index directory for the page and then changing focus to the directory created above so that files can be saved within it:

```
let "index++"
pad='printf "%05d" $index'
mkdir $DATE-$pad
cd $DATE-$pad
```

Wget then takes a screenshot of the page using the programs `wkhtmltopdf` and `wkhtmltoimage` (Truelsen and Kulkarni 2015). It does so using `xvfb-run`, which is an X virtual framebuffer (Wiggins 2015) that allows the program to run without a graphical user interface, making it exceptionally useful because the program can be run without a monitor, or "headless," on an external server. For a long-running, low-maintenance program, this is essential. The two specific programs, `wkhtmltopdf` and `wkhtmltoimage`, do what their names suggest: the first takes an HTML page and turns it into a PDF document, and the second creates a PNG image (screenshot of the site). As we will discuss below, these features proved to be very helpful to researchers. The commands for these programs look like this:

```
/usr/bin/xvfb-run -a -s "--screen 0 1280x1024x24" /usr/bin/wkhtmltopdf --dpi 200 --page-size Letter --custom-header 'User-Agent' 'User-Agent Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; de-de) AppleWebKit/533.19.4 (KHTML, like Gecko) Version/5.0.3 Safari/533.19.4' "$line" $pad-$DATE.pdf
```

```
/usr/bin/xvfb-run -a -s "--screen 0 1280x1024x24" /usr/local/bin/wkhtmltoimage --use-xserver --custom-header 'User-Agent' 'User-Agent Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_6; de-de) AppleWebKit/533.19.4 (KHTML, like Gecko) Version/5.0.3 Safari/533.19.4' "$line" tmp.png /usr/bin/pngcrush tmp.png $pad-$DATE.png rm tmp.png
```

Once the screenshots of the website have been generated, the script moves on to its next task: creating a WARC file. Using Wget flags, the script creates a WARC of the page, saving HTML/CSS documents with proper extensions, capturing all the images and other components needed to display the page, and transforming links in the downloaded HTML or CSS files point to local files (so that on viewing the web page on a local device, the browser does not ascend to the parent directory). All this is done while waiting 1.5 seconds between retrievals, to minimize detrimental harm to the hosting server and to reduce the risk of being flagged as a program with malicious intent (Free Software Foundation 2015). The open-source guide *The Programming Historian* has more information on some of these commands as well as lessons that illustrate how the coding works (Milligan 2012). The full command is as follows:

```
/usr/local/bin/wget --adjust-extension --page-requisites --convert-links --no-parent --random-wait --warc-file=$pad-$DATE "$line"
```

Finally, Wget zips up the page's directory and writes it to a log file. This log file allows users to reconstruct what the crawler did and identify any problems. These commands finish off the script:

```
cd $HOME/FDA_$(DATE)
zip -r $DATE-$pad.zip $DATE-$pad
rm -rf $DATE-$pad
echo "$(date) - $line archived" >> /var/log/daleascii.log
```

Once a given crawl is run, users can take a look at a given SIP. If a crawl is unzipped, the directory will have an index folder that corresponds to the line number in the seed list. A given page's folder will contain a PDF, PNG, WARC, and page directory containing all requests. See [Figure 1](#) for view of directory. See [Figure 2](#) for a more in-depth view of the page directory.

Figure 1: View of directory

```
├── 00001
│   ├── 00001.pdf
│   ├── 00001.png
│   ├── 00001.warc.gz
│   └── web.archive.org
├── 00002
│   ├── 00002.pdf
│   ├── 00002.png
│   ├── 00002.warc.gz
│   └── www.change.org
├── 00003
│   ├── 00003.pdf
│   ├── 00003.png
│   ├── 00003.warc.gz
│   └── leiterreports.typepad.com
├── 00004
│   ├── 00004.pdf
│   ├── 00004.png
│   ├── 00004.warc.gz
│   └── pinboard.in
├── 00005
│   ├── 00005.pdf
│   ├── 00005.png
│   ├── 00005.warc.gz
│   └── chronicle.com
├── 00006
│   ├── 00006.pdf
│   ├── 00006.png
│   ├── 00006.warc.gz
│   └── www.timeshighereducation.co.uk
├── 00007
│   ├── 00007.pdf
│   ├── 00007.png
│   └── 00007.warc.gz
├── 00008
│   ├── 00008.pdf
│   ├── 00008.png
│   ├── 00008.warc.gz
│   └── blogs.princeton.edu
├── 00009
│   ├── 00009.pdf
│   ├── 00009.png
│   ├── 00009.warc.gz
│   └── lawprofessors.typepad.com
├── 00010
│   ├── 00010.pdf
│   ├── 00010.png
│   ├── 00010.warc.gz
│   └── leiterreports.typepad.com
├── 00011
│   ├── 00011.pdf
│   ├── 00011.png
│   ├── 00011.warc.gz
│   └── capalibrarians.org
├── 00012
│   ├── 00012.pdf
│   └── 00012.png
```

```

├── 00012.png
├── 00012.warc.gz
├── plggta.org
├── 00013
├── 00013.pdf
├── 00013.png
├── 00013.warc.gz
├── samtrosow.wordpress.com
├── 00014
├── 00014.pdf
├── 00014.png
├── 00014.warc.gz
├── blogs.princeton.edu
├── 00015
├── 00015.pdf
│   ├── 00015.png
│   ├── 00015.warc.gz
│   └── leiterreports.typepad.com

```

29 directories, 45 files

Figure 2: View of the page directory

```

├── 00012
│   ├── 00012.pdf
│   ├── 00012.png
│   ├── 00012.warc.gz
│   ├── plggta.org
│   ├── archives
│   │   └── 149.html
│   ├── robots.txt
│   ├── wp-content
│   ├── plugins
│   │   ├── contact-form-7
│   │   │   ├── includes
│   │   │   │   ├── css
│   │   │   │   │   └── styles.css?ver=3.3.3.css
│   │   │   │   └── js
│   │   │   │       ├── jquery.form.min.js?ver=3.25.0-2013.01.18
│   │   │   │       └── scripts.js?ver=3.3.3
│   │   ├── google-analyticator
│   │   │   └── external-tracking.min.js?ver=6.4.3
│   │   ├── jetpack
│   │   │   ├── modules
│   │   │   ├── widgets
│   │   │   │   └── widgets.css?ver=20121003.css
│   │   │   └── wpgroho.js?ver=3.5.1
│   ├── themes
│   └── wpbootstrap
│       ├── css
│       │   └── bootstrap.css
│       ├── js
│       │   ├── application.js
│       │   └── google-code-prettify
│       │       ├── prettify.css
│       │       └── prettify.js
│       └── style.css

```

18 directories, 16 files

Next we will look at how the team transformed the captured web pages (SIPs) into AIPs and DIPs.

Islandora solution pack Web ARChive: AIP/DIP dissemination

Once the the PLG-GTA team collected the material, the next question was how to make it accessible and of use to a researcher community. Capturing web pages is only the first step in the process of documenting online history. Without a robust preservation system, accompanied by useful and accessible proxies that can be consulted by researchers, the project team's activities would have been well intentioned but ultimately of little value.

Continuing to view this project through the OAI lens of submission, archival, and dissemination information packages (SIPs, AIPs, and DIPs, respectively), we provided the repository with the SIP created above. The SIP contained three data streams for each object we created: a WARC file, a PDF file, and a PNG file. To handle the transfer, we employed the Islandora Web ARChive Solution Pack. The Web ARChive solution pack provides the means to preserve and disseminate web archives in Islandora. Prior to the creation of this module, not solution existed for the preservation and dissemination of web archives in institutional repository platforms such as Islandora, Project Hydra, or DSpace. There are two options for getting content into Islandora with the solution packs. The first is to add an individual object via Islandora XML Forms ([Banks 2015](#)); the second, to batch ingest with Islandora Batch ([Vessey 2015](#)). The team used Islandora Batch to add over 200,000 distinct objects to the #freedaleasky collection.

Upon ingest, data streams are added to the object via derivative creation and utility modules such as Islandora FITS ([Ruest 2015b](#)) and Islandora Checksum ([Ruest 2015a](#)). Islandora FITS uses the File Information Tool Set (FITS) to complete file identification and characterization. Islandora Checksum creates a checksum of the object that can be checked later for file integrity (fixity). All of the above is driven by the way the object is modelled ([Ruest 2015c](#)) in the Web ARChive solution pack. See [Figure 3](#) for a diagram of a WARC file.

Each Web ARChive object can have up to eleven data streams, including the three data streams created during the SIP phase. The solution pack creates display derivatives—a thumbnail JPG and medium-sized JPG—of the PNG screenshot created in the SIP phase, as well as WARC derivatives using the Internet Archive's WARC Tools ([Internet Archive 2015b](#)). These derivatives include a CSV file made using `warcindex.py` that has an index of all the files in a given WARC (WARC_CSV), as well as a filtered WARC `WARC_FILTERED` made using `warcfilter.py`. This is a WARC file stripped down as much as possible to the text, and it is used only for search indexing and keyword searching.

Figure 3: Diagram of a WARC file

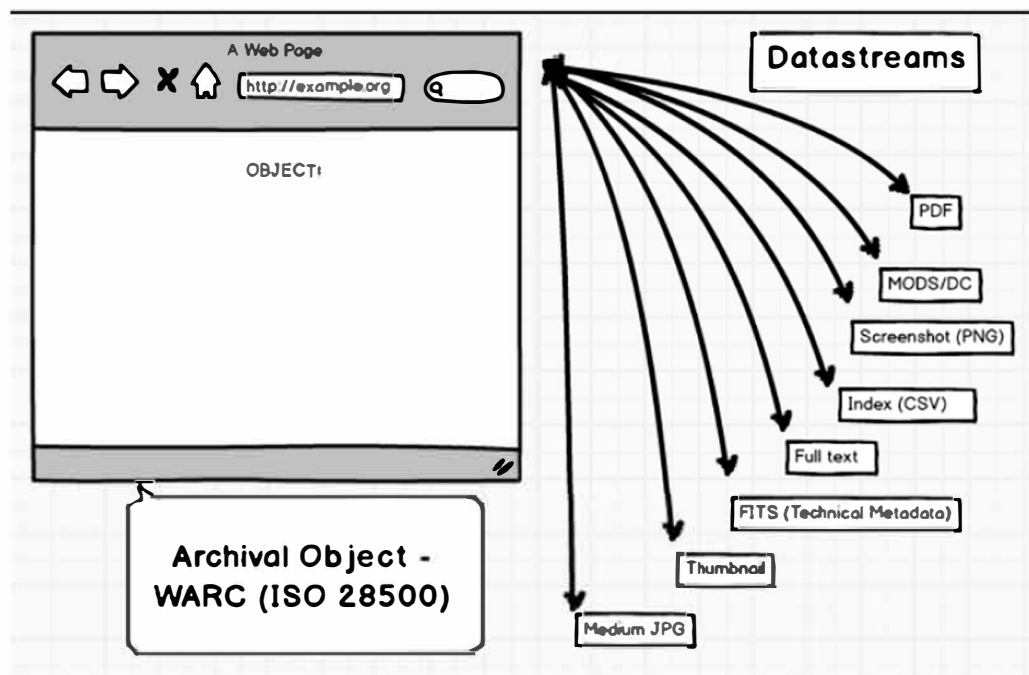
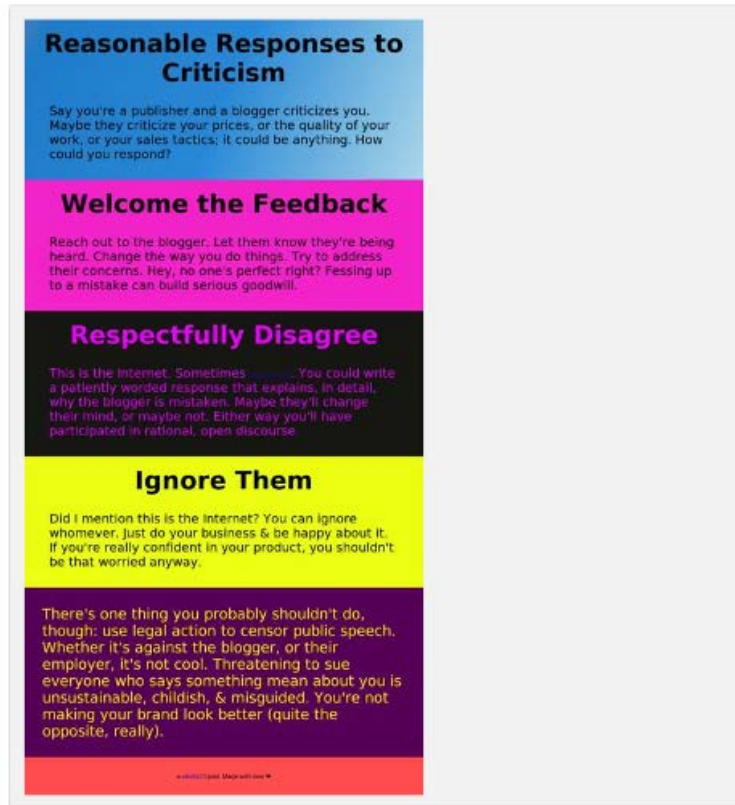


Figure 4: Screenshot of a DIP (Dissemination Information Package) of a WARC file

00190-2013_04_05


[Details](#)

Download

- Warc: 00190-2013_04_05.warc
- PDF: 00190-2013_04_05.pdf
- Screenshot: 00190-2013_04_05.png

In collections

- [Reasonable Responses to Criticism](#)

The DIP is a JPG and thumbnail of the captured website (if supplied) and download links to the WARC, PDF, WARC_CSV, screenshot, and descriptive metadata. An example of a DIP can be seen in [Figure 4](#).

Here a link to the "archived site" can be supplied in the default descriptive metadata form using the Metadata Object Description Schema (MODS). The suggested usage here is to provide a link to the object in a local instance of the Wayback Machine (a way to generate archived websites for viewing, as in the Internet Archive's original instance at <http://archive.org/web>), if such exists.

```
<mods:location>
```

```
<mods:url displayLabel="Active site">http://yfile.news.yorku.ca/</mods:url>
```

```
<mods:url displayLabel="Archived site">http://digital.library.yorku.ca/wayback/20140713/http://yfile.news.yorku.ca/</mods:url></mods:location>
```

The solution pack used in conjunction with the Islandora Collection Solution Pack provides basic parent-child relationship structure. It allows us to organize the archive by seed URLs (parent) and individual crawls (children). The [figure 5](#) below shows the parent:

Figure 5: Parent folders of seed URLs containing individual crawls of sites

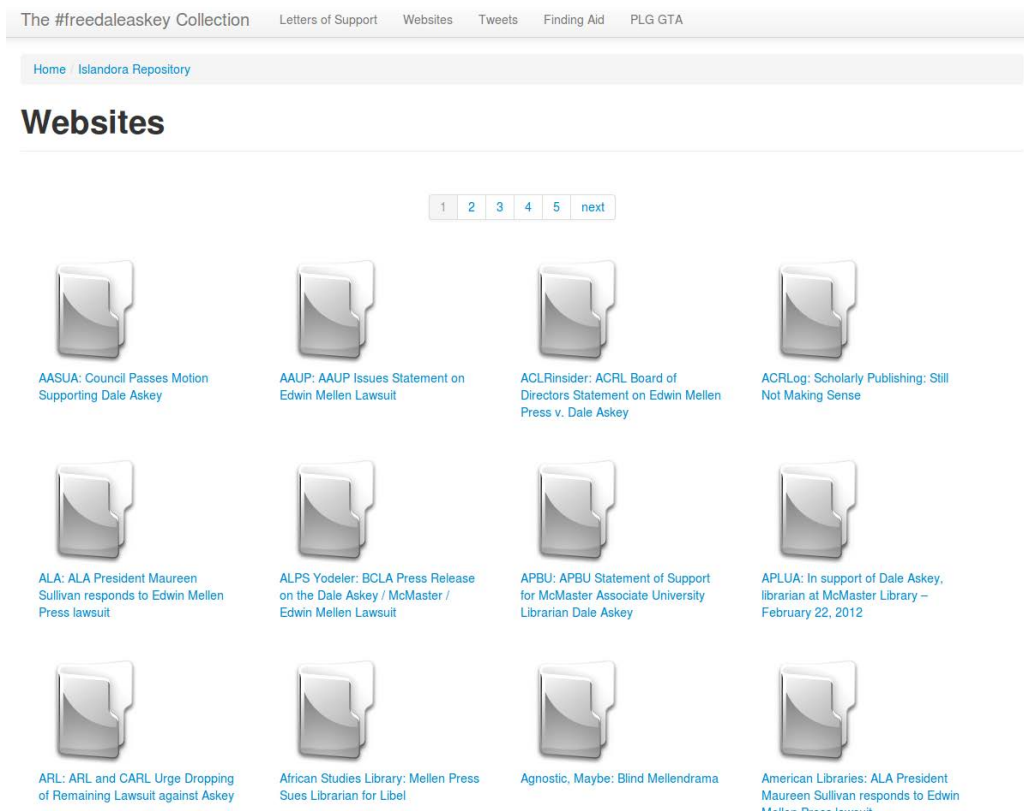
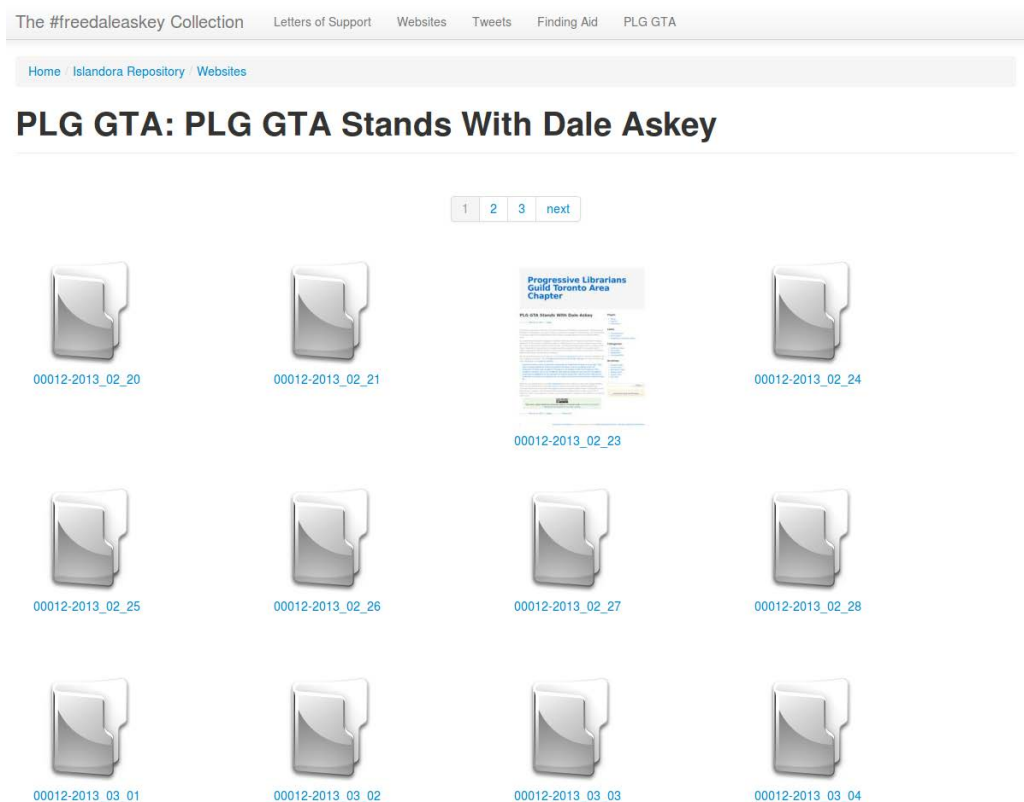


Figure 6 shows the children (individual crawls with standard file names containing date of capture) within specific directories (folders):

Figure 6: Individual crawls of web page "PLG GTA stands with Dale Askey." Note date of crawl is part of file name.



Arrangement

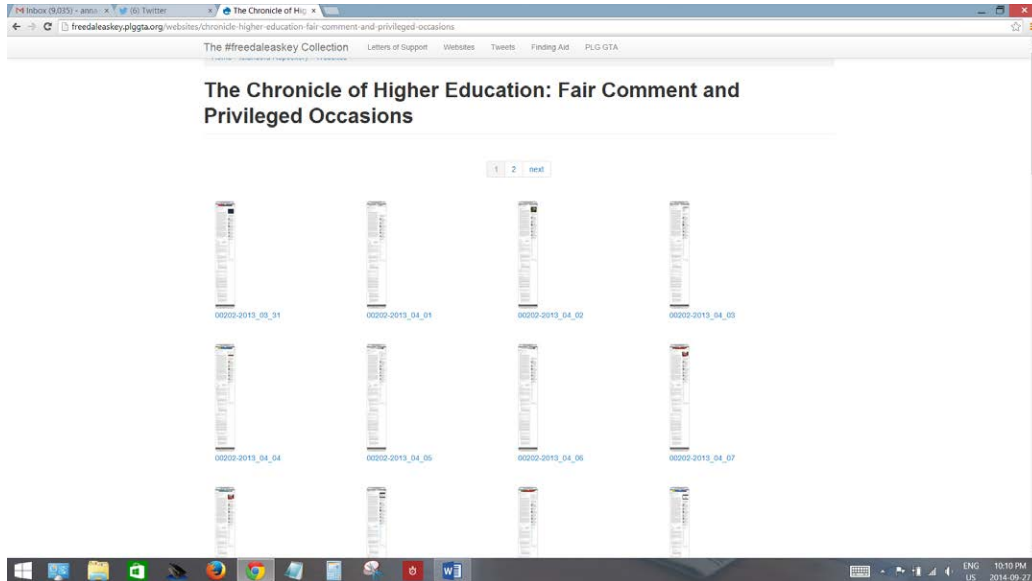
There were two major ways the PLG-GTA team arranged the information. First, the team created the repository itself (freedaleaskey.plggta.org), containing hundreds of individually crawled WARC files for hundreds of distinct websites. The second

(archives.plggta.org) constituted a traditional intellectual arrangement of the collection. This was constructed after the project was well underway, with the purpose of testing the strength and vitality of traditional methods, such as a finding aid, of providing the contents with an interpretive overlay of contextual information.

Islandora repository

The collection is organized by seeds and crawls. Seeds represent each crawled page, drawing from the list of websites we aimed to collect. Each seed is thus an Islandora collection object with its own descriptive metadata. That collection object contains a number of crawls, which are organized by date. You can see an example ([Pullum 2013](#)) in [Figure 7](#).

Figure 7: Example of individual crawls of a blog post from The Chronicle of Higher Education



If you click on the PNG element of the DIP, you will get a long scrolled captured image of the full website, taken on a specific date. See [Figure 8](#).

Figure 8: Example of a captured blog post by Geoffrey Pullum in "Linga Franca" of The Chronicle of Higher Education. Note that some elements, such as the flash advertising on the right does not render correctly.

[Premium Content](#)

[Log In](#)
[Create a Free Account](#)
[Subscribe Now](#)

THE CHRONICLE

of Higher Education

Sunday, March 31, 2013 [Subscribe Today](#)

[HOME](#)
[NEWS](#)
[OPINION & IDEAS](#)
[FACTS & FIGURES](#)
[BLOGS](#)
[JOBS](#)
[ADVICE](#)
[FORUMS](#)
[EVENTS](#)
[STORE](#)



THINK BIBLICALLY. ABOUT EVERYTHING.

→ OPEN BIOLA.edu

Home > Blogs > Lingua Franca

LINGUA FRANCA

Language and writing in academe.

Previous
→ OK, the Gentle Giant
Next
Guys and ... ? →

Fair Comment and Privileged Occasions

March 26, 2013, 12:01 am
By Geoffrey Pullum

I've been interested in the linguistic aspects of defamation law for many years. Delving into the history of libel and slander uncovers all sorts of strange facts. Some are discussed in Chapters 12 and 13 of my book *The Great Eskimo Vocabulary Hoax*, among them a case of a linguistics book that was blocked from publication because lawyers advised that the invented example sentences might be grounds for a libel action.

Under English case law, you can be sued (perhaps even successfully) for the content of interrogatives and imperatives as well as declaratives; for what is presupposed or implied as well as what is said; for statements you don't yourself regard as defamatory; and even for words of praise if a reasonable person would think you were ironically implying something defamatory.

But there are defenses. Justifiable assertion of a provably true claim will normally not be subject to a successful action. Nor will "fair comment" (nonmalicious expression of opinion). Impartial attempts at informing the public on a matter of general interest after diligently seeking to establish the truth is normally defensible. And above all, there is privilege.

Not just utterances in Parliament or in court are privileged. It is normally a good defense that a statement was made in the course of doing a job that




About This Blog

Posts on Lingua Franca present the views of their authors. They do not represent the position of the editors, nor does posting here imply any endorsement by *The Chronicle*.

Questions? Ideas? You can reach us at linguafranca@chronicle.com.

Lingua Franca Bloggers



Anne Curzan
is a professor of English at the University of Michigan, where she also holds appointments in the linguistics department and the

Storage and organization using Fedora commons & Islandora

Quality analysis of the crawl is clear from the visualization of the captured WARC file (i.e., the PNG or JPEG image). At a glance, we can see that the site is being crawled correctly; if we were receiving 404 errors or other codes, they would appear in the screenshots. While not a replacement for close reading, the quick mosaic of images gave the crawl operator an easy way to pick up major changes in the configuration of the website.

Each web archive object has a basic descriptive metadata data stream which captures basic metadata regarding date of capture, size, and unique identifier. In terms of access and reuse, each web archive object has a URI or Uniform Resource Identifier, along with its derivatives. By default, the WARC files themselves are available to download. Preservation is dependent upon the policies of the repository and institution. In the case of York University Libraries, a preservation action plan for web archives and the suite of Islandora preservation modules cover the basic processes (checksum, checksum checker, FITS and PREMIS generation) for preservation and public access.

Intellectual arrangement and creation of a traditional archival finding aid

The PLG-GTA team had captured, preserved, and placed in a digital repository a significant collection of websites in its efforts to document events around #freedaleasky. This wealth of resources had been crawled and preserved in a structured environment. Researchers with access to the collection could find the sites that interested them if they knew what they were looking for or if they had the time and impetus to browse through each folder. But there was no map to help an unfamiliar researcher navigate. The PLG-GTA team thought creating a traditional archival finding aid might provide much-needed intellectual framing, contextual information and authority to what had been captured.

This finding aid lent rhetorical power to our enterprise. Just as the collection of published materials by state institutions tends to lend additional authority to individual works, and just as libraries and archival institutions participate in the construction, maintenance, and posterity of literary canons as well as national and local narratives and histories, the team consciously employed the documentary form of a finding aid to lend authenticity and authority to the stories selected for preservation. In this manner, the team hoped to make transparent any bias and expose the nature of the tools employed. There was also an expectation that the collection of these preserved documents might in some way serve Askey's legal team, so a finding aid to facilitate navigation of the files was deemed to be useful.

Employing OS descriptive software

The PLG-GTA team installed an instance of ICA-AtoM (now known simply as AtoM, or Access to Memory), an open-source descriptive

software system created specifically for archival institutions ([Artefactual Systems 2015](#)). Using it, one team member constructed a logical series system based on form and function. For example, websites that were letters of support from other universities and unions went into one series. The team member subsequently created file-level descriptions based on the individual sites that had been collected.

Drawing on various standards, including Library and Archives Canada's AMICUS database ([2004](#)), OCLC's VIAF for authority control ([2015](#)), this intellectual arrangement allowed the team to create authority records (structured data that is able to differentiate between particular individuals) for authors, anonymous actors, and bloggers who actively participated in the debate, records that could serve as pathfinders to areas of interest or "hot spots" of discussion that team members themselves had identified while going about the work of capturing the sites of discussion.

Providing file-level descriptions also offered a way to document and make transparent the rationale, methodology, and objectives of the PLG-GTA. The approach conceptualized the "object" described as the physical blog post, notably the multiple WARC files captured over time. This way the focus of description was less the written content of the post and the discussion that ensued than when the site came to the attention of the project team, when it was crawled, and any issues that interrupted or disrupted the crawl (failed captures, problematic site structures, etc.).

Reflection on archival practice

Some of the more traditional approaches to archival arrangement and description were disrupted by the nature of #freedaleaskey activities. An example was the discussions among the PLG-GTA team about how to structure the authority records/creator fields and dates and extents. When things really did not seem to work or became uncomfortable to our professional sensibilities, we took it as an indication that we, as archivists, might need to reconsider our approach. The archivist Anna St. Onge, for example, would have preferred a chronological arrangement but the functionality of Islandora was not at its current level, making it difficult to easily check for the most recent additions to the repository. Instead, an alphabetical arrangement was established. As well, archival descriptive standards of the time limited our ability to provide accurate measurement of materials in the collection. How does one measure the physical extent of scraped and archived websites, for example? How many meters of space (or megabytes?) do a thousand tweets occupy?

A historian in the archive

Capturing and preserving the daily iterations of the websites was a worthwhile and useful exercise: it provided a mass of documentation that could be drawn upon by Askey's defence team should any legal proceedings result from the defamation suit. For the individuals "in the know," the records were accessible and could be navigated with relative ease. But was the collection accessible and useful to people not immediately familiar with the Dale Askey case? In the same vein, creation of an intellectual arrangement of the captured websites and public statements proved an interesting practice for the archivist, but it remained to be seen if this tool would provide researchers with useful pathways to navigate and identify areas of interest for further research. The team turned to a historian to get some feedback about the ability of their collecting efforts to generate usual scholarship.

The #freedaleaskey collection was, in many ways, ideal from the perspective of a practising historian because it was compiled from daily scrapes of the Internet and drew upon a substantial corpus of dated material. Daily snapshots allow the historian to trace changes over time, to study the changing relative frequencies of terms, topics, and discussions, and to compare the information to other daily sources. Historians who have used other web archives have run into trouble dating them: if a scrape is only carried out once a year, for example, we may know that a page or change occurred during that time, but extracting precise date information is tricky. One also encounters scenarios in which it is not possible to determine whether an event was an isolated phenomenon or a reflection of broader trends. The #freedaleaskey archive has the added benefit of allowing the historian to reasonably establish that changes occurred on a given day. In this section, we will explore what a user can learn from the #freedaleaskey archives.

In order to effectively work with and manipulate large web archives, source transformations are necessary. Ian Milligan works with web archives, exploring how historians can fruitfully examine the large quantities of information that they encapsulate. To do so, he works largely with textual content. While images are rich sources, it is easier to process text to give a sense of how events evolved over time.

Extracting text from web archives is not a straightforward undertaking. To do so in this case, we employed the WARC Tools kit. An older version of that tool set, hosted on GitHub at <https://github.com/ianmilligan1/Historian-WARC-1/tree/master/WARC/warc-tools-mandel>, turns WARC files into plain text by running each web page through the Lynx web browser. Lynx was one of the first cross-platform web browsers of the early 1990. It is still used today by people with visual difficulties who need to have web pages read to them. Each individual page of the collection went through this process, generating a plain text version with links rendered as endnotes. The downside to this method is that textual content is artificially separated from its graphical content, and screen and other elements that might give context to the text are lost. The advantage of the #freedaleaskey collection is that the researcher can also draw on both the archived web pages and screenshots, which are quicker to access. It is also worth noting that source transformations are nothing new: they are akin to taking notes in an archive, or even taking digital photographs of objects. Scholars do, however, need to be conscious about and make sure to document this process.

One ideal feature of this collection is that each website is preserved in multiple formats, a testament to the complexity of web archiving. Much of the conversation around the Dale Askey case took place in the comment threads of websites. For example, a rather prolific commenter named Thomas Anthony Kelly appeared on a number of forums and blogs discussing the libel case. Characterized by many as a "sock puppet" (i.e., using a pseudonym to seed positive press for the EMP), this commenter contributed almost identical statements across at least four sites captured by the PLG-GTA before being called out by commenter John Jackson ([2013](#)) in a *Scholarly Kitchen* post. For all known sites containing comments by Kelly, see <http://archives.plgqta.org/index.php/thomas-anthony-kelly;isaar>.

In such a charged environment, these comment threads often became more important than the initial posts or articles themselves:

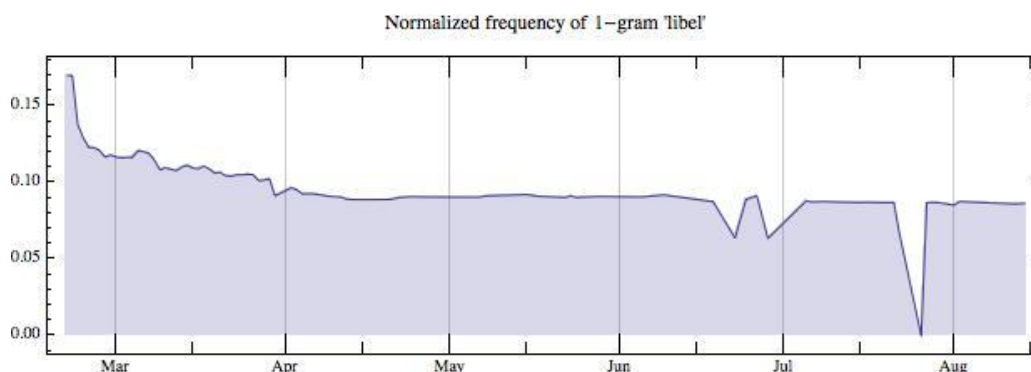
defenders of EMP were occasionally accused of being "sock puppet," or fraudulent online identities created for the singular purpose of defending the publisher; it is also worth remembering that the original litigation was in part against allegedly libellous statements made in the comments thread that Askey did not delete. Yet comments can be very hard to preserve.

Many websites use the popular Disqus platform, which they can host themselves, to help moderate comments, filter spam, and provide an easy-to-use commenting interface (Disqus 2015). Unfortunately, Disqus is hard to archive: it does not immediately load with the site. Looking at the HTML code on a web page from a website like *Insider Higher Ed*, which uses Disqus, one will not see any comments (for a parallel with Adobe Flash, see Ankerson 2012). These were not preserved in the WARC files, or in the PDF versions captured by the crawler. However, they are preserved in the captured screenshots of each page. Although screenshots are not text searchable, they can be run through an Optical Character Recognition (OCR) program to facilitate their study using digital methodologies. Without screenshots, the comments might have been lost to historians forever.

One of the first popular articles written about the libel case, Colleen Flaherty's *Insider Higher Ed* post "Price of a bad review" (2013) illustrates the importance of the comments. In many ways, Flaherty's blog post began the online conversation: in the comments appeared links to the original allegedly libellous post, links to statements from McMaster University as they emerged, links to crucial pieces of litigation, and personal recollections of earlier libel cases. For a few days, "Price of a bad review" served as a community hub for discussions around the case. Comments are crucial ingredients in a web archive of this story. Yet the WARC version and PDF version of the captured *Insider Higher Ed* site do not contain any comments; the screenshot does. This fact is evident from a comparison of the different versions. You can see this for yourself by comparing Flaherty's post (see the screenshot PNG at <http://freedaleaskey.plggta.org/websites/00017-20130306>) with the PDF rendering of the WARC (http://freedaleaskey.plggta.org/islandora/object/plggta%3A9267/datastream/PDF/00017-2013_03_06). These are the smaller details that archivists and historians need to make transparent as we work with these often dynamic collections.

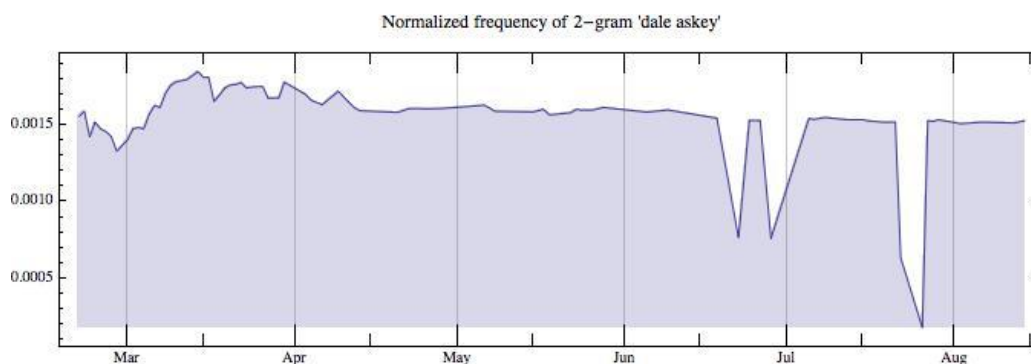
But what can a historian do with all of this textual material? The easiest starting point is simple word frequency over time. Supplied with the date data, n-gram analysis is a useful tool for this purpose (similar to the Google Books n-gram viewer, available at <https://books.google.com/ngrams>). The temporal nature of the #freedaleaskey archive lends itself well to this type of analysis. Standard web archives, such as those available via the Internet Archive's Wayback Machine, have an uneven sampling practice, whereas this collection took a snapshot every day. In Figure 9 is an n-gram for the word *libel*:

Figure 9: N-gram of the normalized frequency of the term *libel* in the collection



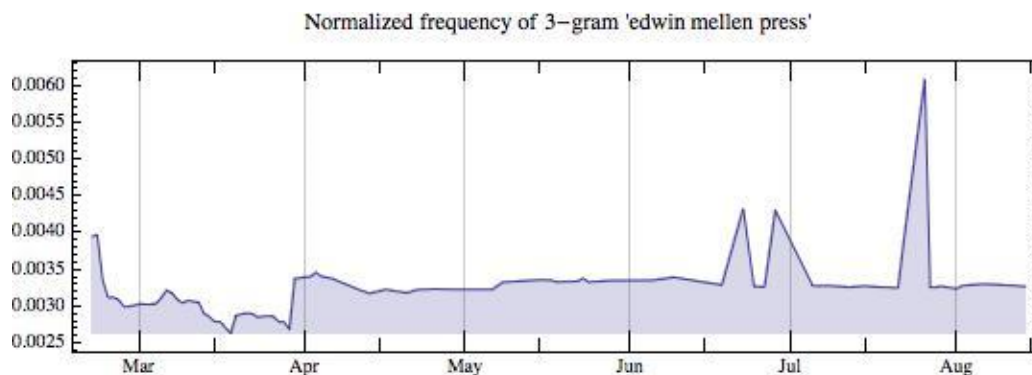
Here we observe that while initial coverage and discussion of the case focused on libel, the word's prevalence tapered off and remained quite stable after early April. Much of this stability reflects a daily scrape of many websites that did not change over time. There is a similar pattern in Figure 10, an n-gram search for *dale askey*:

Figure 10: N-gram of the normalized frequency of the term *dale askey* in the collection



Again, we note discussion throughout March 2013, but a trailing off by the middle of April 2013: conversation during this period shifted away from Askey and discussions of libel toward a broader conversation about scholarly publishing and the

Figure 11: N-gram of the normalized frequency of the term *edwin mellen press* in the collection merits of the Edwin Mellen Press itself.

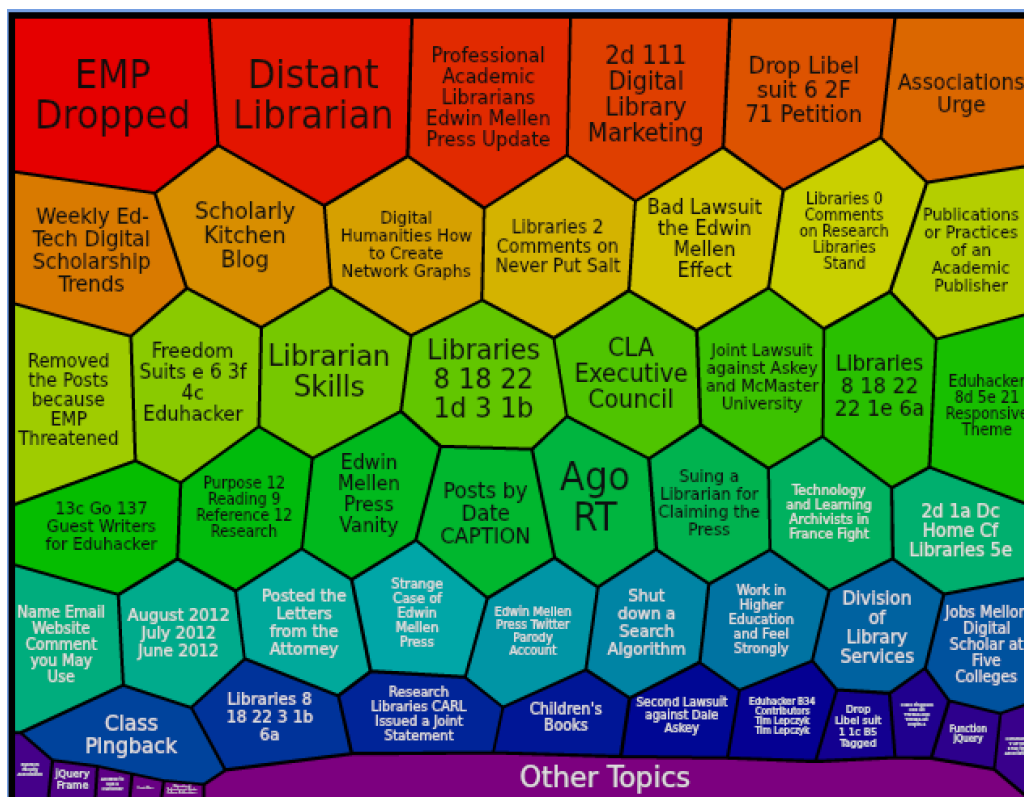


N-gram searches let us trace the evolution of the Dale Askey case. There is value in a mass exploration of WARC files that cannot be derived from close reading of each individual article. Yet these graphs also require some technical knowledge to parse: note the two peaks in late June 2013 in [Figure 11](#) for searches for *edwin mellen press*, and the large peak in late July 2013 matched by an inverse peak in the first figure ([Figure 10](#)). These are technical artefacts of a failed crawl of the sites that represent a gap in the data set rather than a dramatic change in the debate.

One downside, however, of n-grams is that researchers must know what they are looking for. In its implementation, the n-gram is not very fuzzy: a search for *Edwin Mellen Press* produces matches only to that 3-gram (see [Figure 11](#)), not to *Mellen*, *Mellen Press*, or even just *the press*. This issue might be mitigated through more complicated or inclusive search terms, but in general one also wants to find which concepts appear close together. There are two methods to facilitate: the first is clustering search; the second, topic modelling or Latent Dirichlet Allocation (LDA).

Clustering search can be conducted using off-the-shelf research tools. By taking all the text files in the #freedaleaskey collection, ingesting them into Apache Solr (an open-source search engine), and then interpreting them using the Carrot² workbench ([Apache Software Foundation 2015](#); [Weiss and Osinki 2015](#)), we can graphically visualize search queries, as in the following example of *edwin mellen* ([Figure 12](#)):

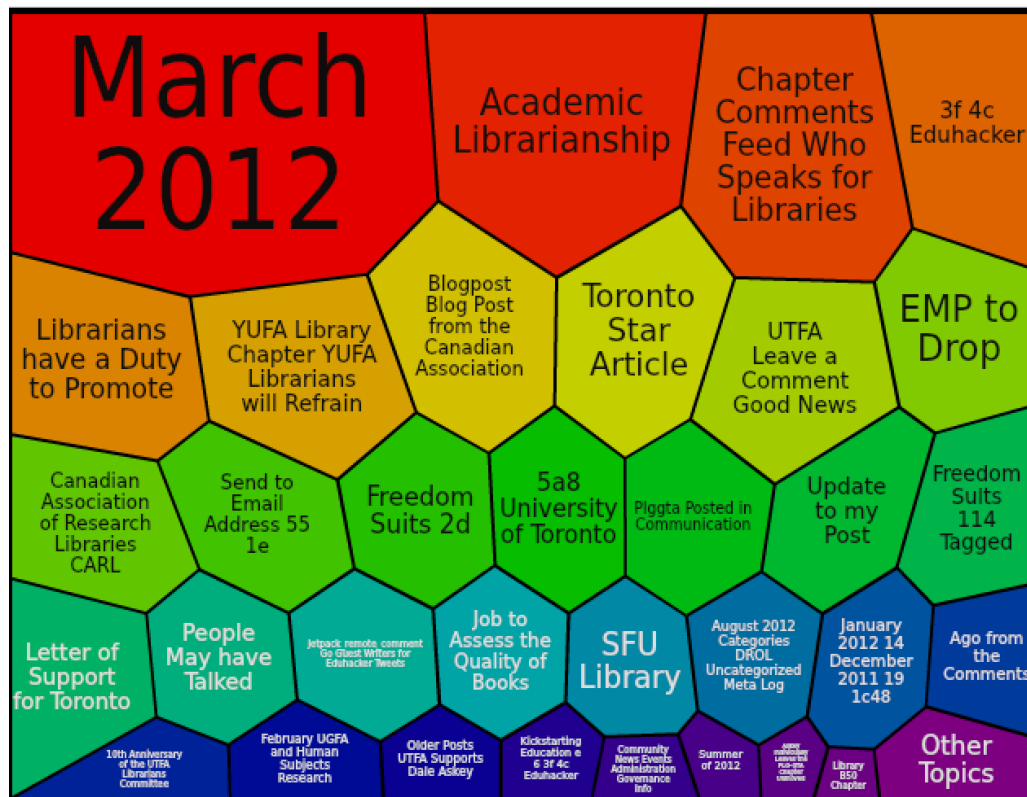
Figure 12: Cluster display of search term *edwin mellen* using Carrot²



Although this visualization might seem quite complex, users encounter clustering often on the Web, even if it may not be presented under the name Apache Solr or Carrot². A Google News search for *hockey*, for example, will produce a headline and a few suggested stories, but underneath that will appear "explore more," with dozens or even hundreds of other articles. Google has clustered the news articles together, and only shows a few top-ranked and representative articles to the reader in the same manner [Figure 12](#) does.

Clustering provides a broader perspective on related elements in the #freedaleaskey collection: *EMP* (the abbreviation for Edwin Mellen Press) is clustered with the term *Dropped*, so a click on that tab will bring up stories around the dropping of the lawsuit. In the same manner, clusters including terms such as *petition*, *removed the posts because EMP threatened* and *lawsuit* will lead to discussions focusing on specific topics that orbited the Askey case. Although this process is not perfect—there is some

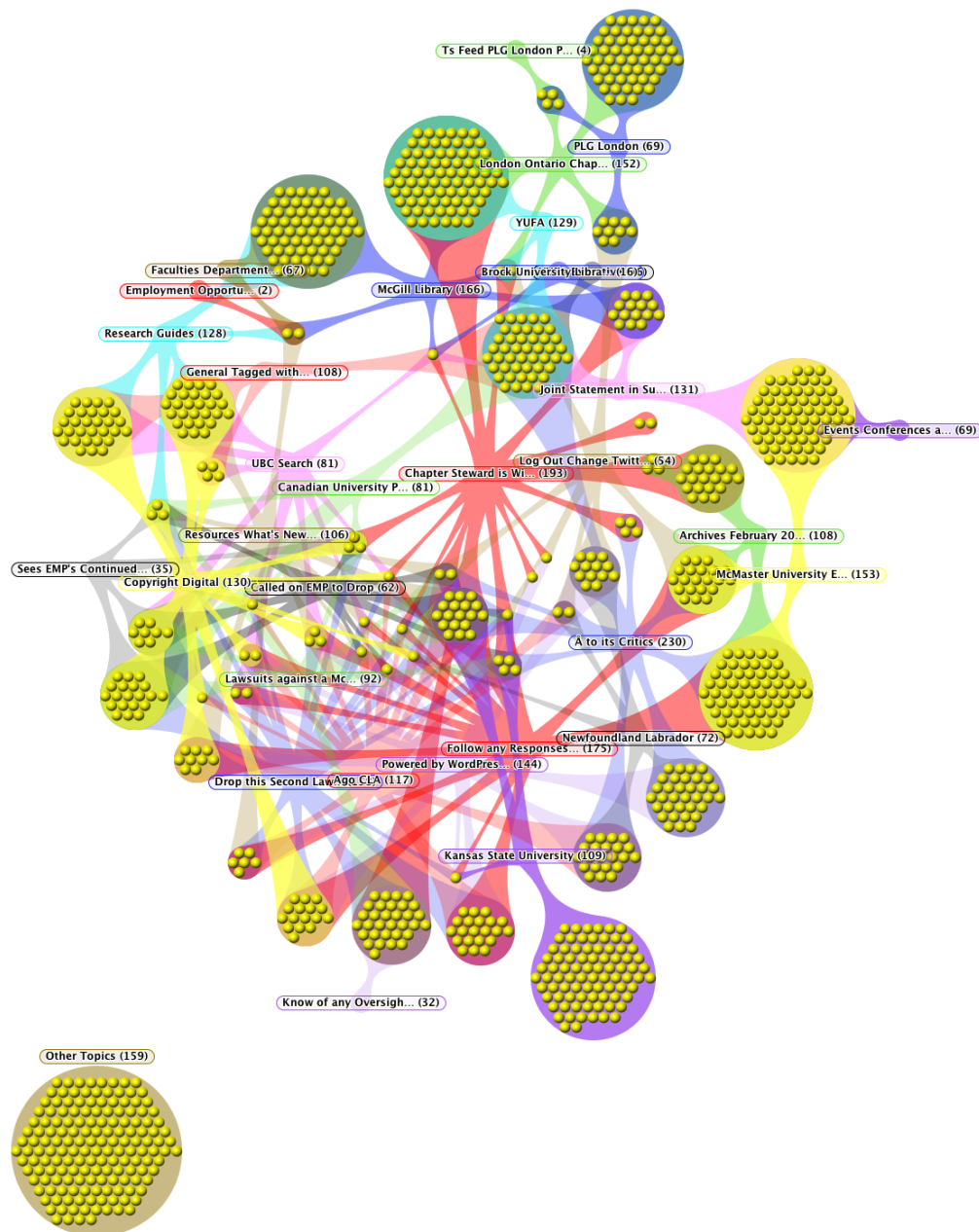
Figure 13: Cluster display of search term librarians using Carrot²



messy data in there—the broad contours of the story are now visible. For a historian, this presents an alluring non-chronological, thematic approach to getting a distant measure of a dataset.

Carrot² is a quick way to graphically make sense of a large web collection. There are other visualizations (Figure 14) which show connections between clusters. Websites might belong to more than one cluster. In the figure below, generated using Aduna Cluster Map software, each circle represents a website, connected to a "label" (string of words) that represents what the websites are about. Websites connected to two labels span them (Aduna 2015).

Figure 14: Scatter topic graph pertaining to Dale Askey using Aduna Cluster Map software.



In the above graph, topics which pertain to *Dale Askey* are placed into larger groups: Brock University Librarians, statements, UBC, the Progressive Librarians Guild London and how they relate to each other branch in the upper right; to the left are discussions around copyright and lawsuits; at bottom and to the right, clusters relating to McMaster (Askey's current employer) and Kansas State (his employer when he wrote the two offending blog posts). In some ways, this is a more complicated version of [Figures 12](#) and [13](#), generated with Carrot². This approach lets researchers visualize the textual content and context of these web archives. A click on any of the circles brings the user to the full text of the original website.

The foregoing methods of keyword searching and clustering have a fundamental disadvantage: users need to know what they are looking for. Topic modelling presents a way around this obstacle. The basic idea behind it is that documents are composed of various topics, or subjects, which themselves are composed of the specific words people choose to make their central arguments. Imagine that a historian is writing an article about working-class women as well as male-dominated unions: when she writes about the first topic, she may use words like *women*, *girls*, *femininity*, *shifts*, *differential*, *feminism*, and in the second one, she might use words such as *steward*, *masculinity*, *differential*, *overtime*, and so forth. Topic modelling, implemented using algorithms such as Latent Dirichlet Allocation (LDA), reverse engineers this process (see [Jockers 2013](#); [Graham, Weingart, and Milligan 2012](#)). Again, due to the date-sequenced nature of the data, one is able to trace the rise and fall of various topics in this corpus. There are no magic methods to topic modelling; rather, they require some finessing with respects to the number of topics and the number of iterations retrieved.

With some tweaking, we ran a topic model on the *#freedaleaskey* (inspired by venues, such as [Summers 2015](#); [Owens 2014b](#) who have proposed topic modelling as a potential finding aid. Fifty topics were extracted from the corpus and divided into three categories: descenders, or topics that became less prominent as *#freedaleaskey* progressed; ascenders, or topics that became *more* prominent; and blips, which saw sharp spikes.

Ten topics decreased in frequency over the time period. They are visualized below in [Figures 15](#) and [16](#). Note that there is no one way to read topic models; rather, they are a conversation starter and let researchers know what they might find within a collection. The number on the left-most column refers to the topic itself.

Figure 15: LDA topic descenders

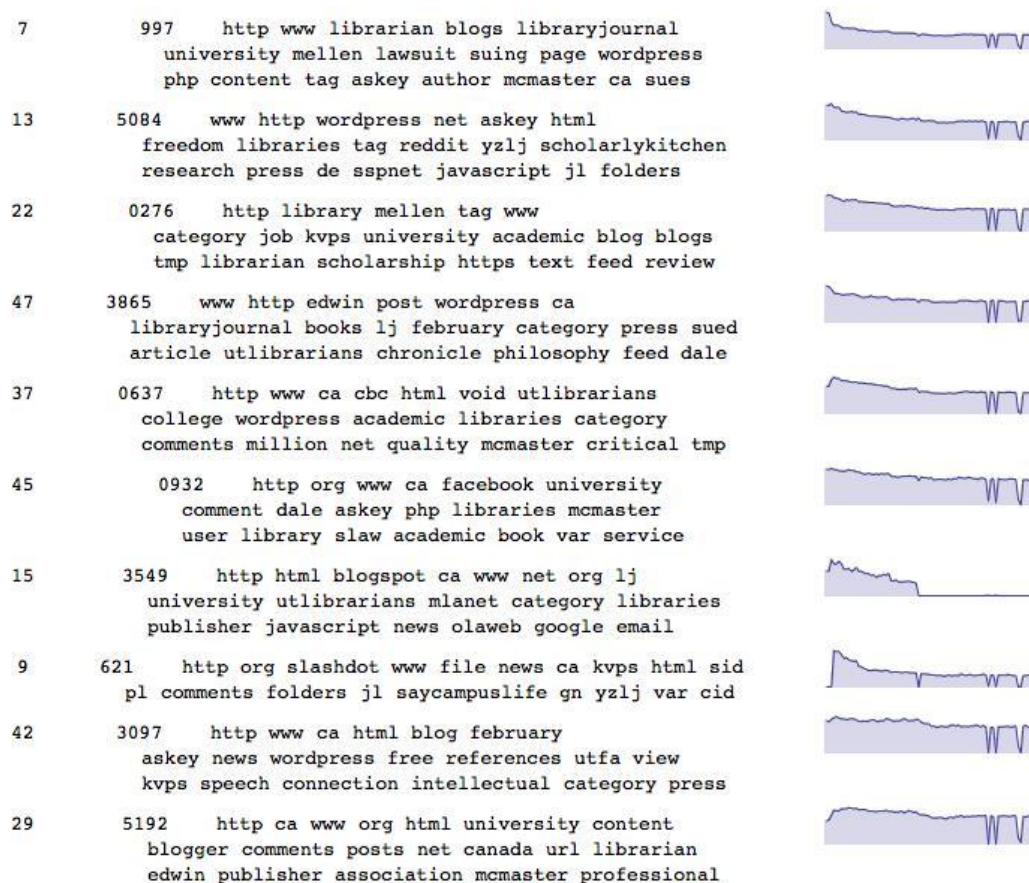
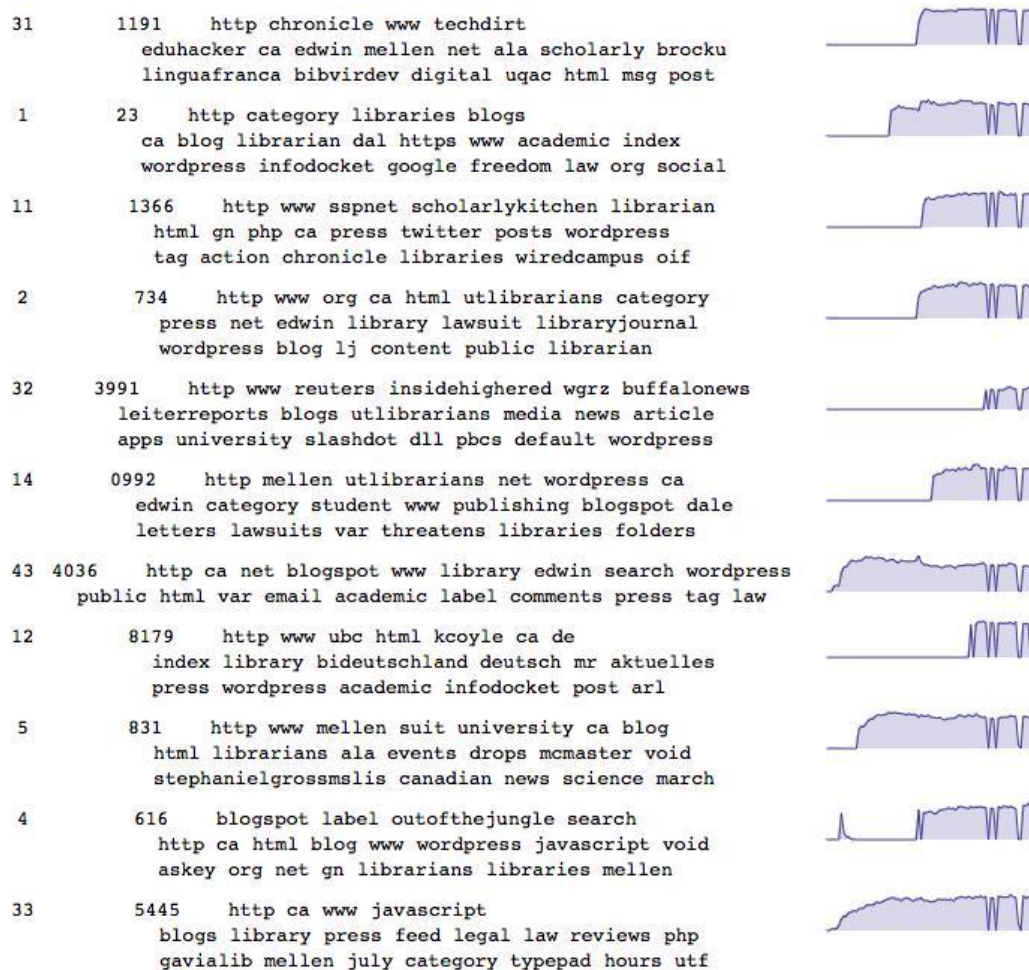


Figure 15 displays all the topics that represented very important articles and initial exposés but generally decreased in attention and significance as other sources came online (again, note that the three hasty reductions to zero in the count are artefacts of the web archiving process). Some topics with steeper slopes reflect the waning attention of larger mainstream sites such as CBC.ca and Slashdot, or the shift by some librarian blogs to less specific and more overarching issues related to the case. Other topics had more shallow decreases; sites that related to McMaster, for example,

Figure 16: LDA topic ascenders



had more staying power.

The "ascenders" are more revealing. [Figure 16](#) illustrates topics that appeared at varying points in the archive, a testament to the story's slow spread and build.

All of these different approaches to visualizing the #freedaleaskey collection help to demonstrate that well-constructed, event-focused and temporally constructed web archives can be useful for historians and other consumers of web archives. Researchers will have an excess of text to deal with in the new web era of scholarship, and these tools help impose order on an otherwise unwieldy amount of information. Given the scope of the #freedaleaskey collection, researchers cannot rely solely on detailed finding aids or subject matter specialists. Topic modelling, keyword searching, and clustering analysis can help bring important themes to the fore. In a straightforward manner, drawing on open-source software, the narrative of an event can be crafted from its online traces. But it depends in the first place on an approach like that of #freedaleaskey to preserve the documentary record: from blog posts, to discussion forums, to news coverage. This form of preservation, access, and textual research should become a model for organizations with a limited budget that need to preserve web data. Although efficient and reliable commercial solutions exist, the authors purposefully sought out solutions that were free to download, were open source in practice, and had a healthy development community. The intent was to demonstrate that even with a limited budget, similar solutions could be built with a modest amount of time, labour, and infrastructure.

Articulating a (best) practice

The #freedaleaskey project was an effort founded and pursued in the spirit of shared professional values, concern for a respected colleague, and a desire to test and practice skills in the aid of a wider social issue: that of free speech and academic freedom in the sphere of online professional blogging and the intersection of scholarly publishing and library collection development. As an open-source project, #freedaleaskey has also contributed back to the communities it drew on. The modified WARC solution pack was returned to the Islandora Foundation and has been incorporated into the latest iteration of the program as a standard solution pack.

The team's efforts were a gesture of solidarity with a colleague who has provided mentorship and leadership within the library community. But it was also hoped the project would signal to the wider community of librarians and archivists the commitment of informal organizations like the PLG-GTA to the shared principles of academic freedom and free speech. We hold that the Internet should be a forum of free and open dialogue without threat of intimidation from parties seeking to muzzle, repress and silence outspoken members of the community or otherwise create a chill on public discourse.

As the #freedaleaskey collection has shown, current events that spark and gather momentum through online social media forums can and should be captured and preserved (in as large a sample as possible) for future scholarly use, social accountability, and the democratic process. Major current events of the past three years in North America as well as more localized community events (from #Ferguson to #GamerGate to #IdleNoMore to #TeamHarpy), have been driven, accelerated, and drawn out in public forums.

The current state of online hosting software does not afford the luxury of waiting for the passage of time to filter out the chaff and deposit the relevant materials into heritage institutions such as libraries, archives, museums, and galleries. In addition to managing those records that naturally find their way into institutions, archivists, as a profession, need to be the advance scouts who capture, preserve, and document—not only for our institutions, but for wider communities, be they local concerns, social circles, professional associations or social justice advocates. This is not a call for librarians and archivists to become citizen journalists or community activists. We are not required to espouse any particular social or political position, but we are obliged, by the professional ethics of libraries and archives, to choose a community to document, preserve, and support. In-depth, focused, precision exercises such as #freedaleaskey can provide local studies that complement, challenge, or add texture to those large-scale web archiving programs underway.

As a collaborative research project, #freedaleaskey also shared and developed expertise that could not have been supported under a traditional single-researcher approach. Anna St.Onge, an archivist, was exposed to tools and techniques employed by her co-authors that have incredible potential for the not-so-distant appraisal and accessioning of born-digital records or similar web archives. It was also instructive to discuss and observe how a researcher went about accessing and constructing meaning out of archival materials. So often, archivists speculate about user behaviour. This was an opportunity to reflect and adjust practice to make material more accessible, without compromising descriptive context and the authenticity of the record.

As a historian, Ian Milligan was delighted to be able to work closely with a librarian and an archivist—to attend their conferences, to e-mail back and forth, and to think more about the theory and work that goes into making collections available. One cannot simply "grab material" and "make an archive." Instead, we need to think consciously about collection procedures, digital preservation, and making archives accessible in a reliable fashion. Historians need to work more closely with librarians and archivists, and learn from them as we take our own first steps into this new digital age.

As an archivist-librarian and developer, Nick Ruest found it extremely helpful to work with a group of highly qualified professionals in the creation of systems for capture and delivery of current-event-based web archives. Above all, the interplay between historian, archivist, librarian and developer roles was an invaluable experience. It inspired and informed the creation of tools for the information professional.

Afterword

Although EMP dropped its defamation suit against McMaster University in March 2013, it continued to pursue Dale Askey in a civil suit seeking \$1.5 million in compensation. On 4 February 2015, Askey announced that Herbert Richardson and EMP had formally dropped both cases against him and that the two parties had reached a legally binding settlement which "mutually releases all parties from any claims" (Fabris 2015). A scanned copy of the court ruling can be found at <https://drive.google.com/file/d/0B6N4V-cokle9SjZxbVpSVGZ6dDBmdks2cjFXakQyRDFRTU9J/view>. "The outcome of this case is essentially a neutral outcome for academic freedom," he said. "Both parties walk away from the matter admitting nothing and resolving nothing" (as quoted in Fabris 2015). The authors hope that despite the neutral result of the story, the #freedaleaskey collection will be used for future study, testing, and instruction by librarians, archivists, and historians alike.

Works Cited / Liste de références

- Aduna. 2015. "Cluster map." *Aduna-software.com*. Accessed February 25. <http://www.aduna-software.com/technology/clustermap>.
- Amendola, Amanda R., to Kent Anderson and Phil Davis. 2013. The #freedaleaskey Collection, 00177-2013_03_30 (screenshot). Letter dated March 13. <http://archives.plggta.org/index.php/scholarly-kitchen-posts-removed-because-weve-received-letters-from-edwin-mellen-press-attorney:isad>.
- Anderson, Kent. 2013. "Posts removed because we've received letters from Edwin Mellen Press' attorney." *The Scholarly Kitchen*, March 29. The #freedaleaskey Collection, 00177-2013_03_30 (screenshot). <http://freedaleaskey.plggta.org/websites/scholarly-kitchen-posts-removed-because-we%E2%80%99ve-received-letters-edwin-mellen-press%E2%80%99-attorney>.
- Ankerson, Megan Sapnar. 2012. "Writing web histories with an eye on the analog past." *New Media Society* 14:384.
- Apache Software Foundation. 2015. Solr (website). Accessed February 25. <http://lucene.apache.org/solr/>.
- Artefactual Systems. 2015. AtoM (website). Accessed February 25. <https://www.accesstomemory.org/en/>.
- Arvidson, Allan, Krister Persson, and Johan Mannerheim. 2000. "The Kulturarw3 project - the Royal Swedish web Archiw3e - an example of 'complete' collection of web pages." Presented at the 66th IFLA Council and General Conference, Jerusalem, Israel, August 13-18. <http://archive.ifla.org/iv/ifla66/papers/154-157e.htm>.
- Askey, Dale. 2010a. "The curious case of the Edwin Mellen Press." *Eintauchen: dive in*, September 22. Wayback Machine. Internet Archive. <http://web.archive.org/web/20110630153231/http://htwkbk.wordpress.com/2010/09/22/the-curious-case-of-edwin-mellen-press/>.
- . 2010b. Comment to Stephen Roberts on Askey, "The curious case of the Edwin Mellen Press." Last modified November 12, 20:42.
- Banks, Nigel (maintainer). 2015. "Islandora/islandora_xml_forms." Accessed February 25. https://github.com/Islandora/islandora_xml_forms.
- Bibliothèque nationale de France. 2006. "Web archiving at BnF." *BnFnews*, September.

<http://www.netpreserve.org/sites/default/files/resources/BnFnews200609.pdf>.

British Library Web Archiving Team. 2014. "The British library collection development policy for websites." The British Library. http://www.bl.uk/aboutus/stratpolprog/digi/webarch/bl_collection_development_policy_v3-0.pdf.

Dewey, Caitlin. 2014. "How web archivists and other digital sleuths are unraveling the mystery of MH17." *Washington Post*, July 21. <http://www.washingtonpost.com/news/the-intersect/wp/2014/07/21/how-web-archivists-and-other-digital-sleuths-are-unraveling-the-mystery-of-mh17/>.

Disqus. 2015. "Supercharging: Below the fold." Accessed February 25. <https://disqus.com/websites/>.

Fabris, Casey. 2015. "Librarian says academic press has settled lingering lawsuit against him." *The Chronicle of Higher Education*, February 5. <http://chronicle.com/blogs/ticker/librarian-says-academic-press-has-settled-lingering-lawsuit-against-him/93413>.

Flaherty, Colleen. 2013. "Price of a bad review." *Inside Higher Ed*, February 8. <http://www.insidehighered.com/news/2013/02/08/academic-press-sues-librarian-raising-issues-academic-freedom#sthash.JG6H4m0l.dpbs> [or captured WARC file at <http://freedaleaskey.plgta.org/websites/inside-higher-ed-price-bad-review/>].

Free Software Foundation, 2015. *GNU Wget 1.16.1 Manual*. Accessed February 25. <http://www.gnu.org/software/wget/manual/wget.html>.

Graham, Shawn, Scott Weingart, and Ian Milligan. 2012. "Getting started with topic modeling and MALLET." *The Programming Historian*, September 2. <http://programminghistorian.org/lessons/topic-modeling-and-mallet>.

International Internet Preservation Consortium. 2015. "Members." Netpreserve.org. Accessed February 25. <http://live.iipc.gotpantheon.com/about-us/members>.

Internet Archive. 2015a. "Frequently asked questions." Internet Archive. Accessed February 25. https://archive.org/about/faqs.php#The_Wayback_Machine.

———. 2015b. "internetarchive/warctools." GitHub. Accessed February 25. <https://github.com/internetarchive/warctools>.

Jackson, John. 2013. Comment on Rick Anderson, "When sellers and buyers disagree—Edwin Mellen Press vs. A critical librarian." *The Scholarly Kitchen*, February 11. Last modified February 26, 13:14. The #freedaleaskey Collection, 00042-2013_03_06 (screenshot). <http://freedaleaskey.plgta.org/websites/scholarly-kitchen-when-sellers-and-buyers-disagree-%E2%80%94-edwin-mellen-press-vs-critical>.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.

Kahle, Brewster. 1996. "Archiving the Internet." *Scientific American*, April 11. <http://www.uibk.ac.at/voeb/texte/kahle.html>.

Kaplan, David E. 2014. "Calling for back up" (radio interview). By Bob Garfield. *On the Media*, March 7. <http://www.onthemedial.org/story/calling-back/>.

Kuny, Terry. 1997. "A digital dark ages? Challenges in the preservation of electronic information." *63rd IFLA Council and General Conference*, August 27. <http://archive.ifla.org/IV/ifla63/63kuny1.pdf>.

Lesk, Michael. 1995. "Preserving digital objects: Recurrent needs and challenges." Lesk.com. <http://www.lesk.com/mlesk/auspres/aus.html>.

Library and Archives Canada. 2004. "What is AMICUS?" Amicus Canadian National Catalogue. Last modified January 1. <http://www.collectionscanada.gc.ca/amicus/006002-122-e.html>.

Milligan, Ian. 2012. "Automated downloading with Wget." *The Programming Historian*, June 27. <http://programminghistorian.org/lessons/automated-downloading-with-wget>.

———. 2014. "Preserving history as it happens: The Internet archive and the Crimean crisis." *ActiveHistory.ca*, March 25. <http://activehistory.ca/2014/03/preserving-history-as-it-happens/>.

National Library of Australia. 2009. "History and achievements." PANDORA: Australia's Web Archive, February 18. <http://pandora.nla.gov.au/historyachievements.html>.

OCLC. 2015. "VIAF: Virtual International Authority File." OCLC.org. Accessed February 15. <http://www.oclc.org/viaf.en.html>.

Owens, Trevor. 2014a. "What do you mean by archive? Genres of usage for digital preservers." *The Signal: Digital Preservation*, February 27. <http://blogs.loc.gov/digitalpreservation/2014/02/what-do-you-mean-by-archive-genres-of-usage-for-digital-preservers/>.

———. 2014b. "Mecha-Archivists: Envisioning the role of software in the future of archives." *TrevorOwens.org*, May 27. <http://www.trevorowens.org/2014/05/mecha-archivists-envisioning-the-role-of-software-in-the-future-of-archives/>.

PLG-GTA (Progressive Librarians Guild Toronto Area Chapter). 2015. "About." Accessed February 25. <http://plgta.org/about>.

Pullum, Geoffrey. 2013. "Fair comment and privileged occasions." *Lingua Franca. The Chronicle of Higher Education*, March 26. The #freedaleaskey Collection, 00202-2013_03_31 (screenshot). <http://freedaleaskey.plgta.org/websites/chronicle-higher-education-fair-comment-and-privileged-occasions>.

- Robinson, Sale. 2013. "ALA joins protest of SLAPP lawsuit brought by publisher against librarian." *Moby Lives*, February 25. <http://www.mhpbooks.com/ala-joins-protest-of-slapp-lawsuit-brought-by-publisher-against-librarian/>.
- Ruest, Nick (maintainer). 2015a. "Islandora/islandora_checksum." Accessed February 25. https://github.com/Islandora/islandora_checksum.
- . 2015b. "Islandora/islandora_fits." Accessed February 25. https://github.com/Islandora/islandora_fits.
- . 2015c. "Islandora/islandora_solution_pack_web_archive." Accessed February 25. https://github.com/Islandora/islandora_solution_pack_web_archive/blob/7.x/xml/islandora_web_archive_ds_composite_model.xml.
- Summers, Ed. 2015. "edsu/fondz." Accessed February 25. <https://github.com/edsu/fondz>.
- Taylor, Nicholas. 2014. "The MH17 crash and selective web archiving." *The Signal: Digital Preservation Blog*, July 28. <http://blogs.loc.gov/digitalpreservation/2014/07/21503/>.
- Theimer, Kate. 2012. "Archives in context and as context." *Journal of Digital Humanities* 1(2). <http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/>.
- Truelsen, Jakob, and Ashish Kulkarni. 2015. "WK<html>Topdf." Accessed February 25. <http://wkhtmltopdf.org/>.
- Vessey, Adam (maintainer). 2015. "Islandora/islandora_batch." Accessed February 25. https://github.com/Islandora/islandora_batch.
- Weiss, David, and Stanislaw Osinki. 2015. Carrot² (website). Accessed February 25. <http://project.carrot2.org/>.
- Wiggins, David P. 2015. "XVFB." X.org Foundation. Accessed February 25. <http://www.x.org/releases/X11R7.6/doc/man/man1/Xvfb.1.xhtml>.
- York University. 2015. "Common records schedule." Information and Privacy Office. Accessed February 25. <http://crs.apps06.yorku.ca/>.

