Archiving the Web and Beyond: A Look at Twitter and Facebook (and some other things too)

July 23, 2014

Benn Joseph
Manuscript Librarian
Northwestern University Library
Outline of discussion

• Digital preservation management (in general).
• How (and why) we archive websites at Northwestern University (and how to make the case).
• Social media archiving and the story of the Twitter archive at LC.
• Behind the scenes: how to get your data out.
• Technology and tools: how do we do it?
• Sustaining the collection: ensuring preservation and understanding the costs.
Digital preservation management
Backups vs. Archival Storage

Backups keep your computer working and files safe

Archival Storage keeps the content accessible for future computers and users
An ‘Archival’ storage would comprise:

• Objects = (files + metadata)
• Storage system may include a variety of formats
• Metadata: identification and description
  – Makes object understandable
  – Unique identifier == traceable in time
Metadata uniquely identifies digital objects

Published online before print
March 18, 2013, doi:
10.1177/0165551513480107

Journal of Information Science March
18, 2013 0165551513480107

Unique Identifiers can also store descriptive data
EX: MFM121587

(10) Patent No.: US 6,539,396 B1
Arranging digital objects

Digital photos tossed in a folder for future archiving, credit: Keri Myers (a volunteer archivist with NDIIPP).
Lots of Copies

“...let us save what remains not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”

-Thomas Jefferson, 1791
How many?

2-6 COPIES!

• At least 2 copies in 2 places
• Consider
  – File size
  – Personally Identifiable Information (PII)
Online Storage

- Cloud and collaborative storage
Dangers! to digital objects

Change and loss – accidental and intentional

These black Write Protect Tabs should be used when you do not want stored information on your disk to be changed or lost accidentally.

Simply cover the Write Enable Notch by taking 1 black tab and wrapping it around the notch.
Dangers! to digital objects

Obsolescence – as technology evolves
Dangers! to digital objects

Inappropriate access – e.g., confidential data
Dangers! to digital objects

Non-compliance – standards and requirements
Dangers! to digital objects

Disasters – emergencies of all kinds
Everyday Protection

• Know where your content is located
  – Onsite and offsite; online and offline (administrative metadata)
• Know who can have access to it
• Know who accesses your secure information
  – For staff, depositors, users (Authentication)
Web Archiving and how we archive websites at NU

- Web Archiving Service
- Product of the California Digital Library
- Subscription-based service
- Offers support
- Mildly customizable

http://was.cdlib.org
WASsup?
A little background on our Web Archiving Service implementation

- 377 GB of data used (1TB available).
- 231 crawled sites.
- Planning began in 2010, we began using in June 2011.
- UA is institutional admin, I am the only project admin and we really only have two projects (one that takes in 99% of what we crawl; the other is dark).
Program of African Studies Newsletter: only being made available online (no longer printed)
Web Archiving Task Force (WATF)
Questions remained...
New library website launched August 2010

Old (top) and new (right) sites
<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPforSSVP/</td>
<td>16-Apr-2008 20:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TheBeat/</td>
<td>28-Oct-2008 20:43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aab/</td>
<td>29-Jul-2011 17:07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>acm/</td>
<td>29-Jul-2011 17:16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aviv/</td>
<td>07-Nov-2005 17:50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>addHarvy/</td>
<td>09-Apr-2006 15:27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alphadelphi/</td>
<td>12-Jan-2006 11:31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anna/</td>
<td>08-Apr-2006 03:09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>amy/</td>
<td>05-Aug-2011 16:31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anna/</td>
<td>06-Nov-2007 18:57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>av/</td>
<td>06-Apr-2007 22:38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>barbershop/</td>
<td>07-Nov-2005 17:50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bills/</td>
<td>27-Mar-2006 20:53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bodenhausen/</td>
<td>07-Nov-2005 17:51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bridge/</td>
<td>07-Nov-2005 17:56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bsg/</td>
<td>07-Nov-2005 17:51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>campkessen/</td>
<td>03-Jan-2007 18:23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>catholics/</td>
<td>07-Nov-2005 17:54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legacy student organization websites
AdreNUline Circus

its a rush.

[enter]
**Events**

- Next meeting: January 2005!

**Goals**

- To promote open standards
- To share knowledge and skills
- To think up a third bullet point

**Contacts**

- Kurt McKee
- Allan Tokuda
Feinberg School of Medicine
Setting up crawl in WAS panel
Departments, Centers & Institutes

Departments
- Anesthesiology
- Cell and Molecular Biology
- Dermatology
- Emergency Medicine
- Family and Community Medicine
- Medical Social Sciences
- Medicine
- Microbiology-Immunology
- Molecular Pharmacology and Biological Chemistry
- Neurological Surgery
- Neurology, Ken and Ruth Davie department
- Obstetrics and Gynecology
- Ophthalmology
- Orthopaedic Surgery
- Otolaryngology—Head and Neck Surgery
- Pathology
- Pediatrics
- Physical Medicine and Rehabilitation

Centers, Institutes, and Programs
- Asher Center for the Study and Treatment of Depressive Disorders
- Biostatistics Collaboration Center
- Bluhm Cardiovascular Institute
  - Buehler Center on Aging, Health & Society
- Center for Behavioral Intervention Technologies
- Center for Comparative Medicine
- Center for Education in Medicine
- Center for Genetic Medicine
- Center for Global Health
- Children’s Memorial Research Center
- Cognitive Neurology and Alzheimer’s Disease Center
- Comprehensive Transplant Center
- Feinberg Cardiovascular Research Institute
- Feinberg Clinical Neuroscience Research Institute
- HIV/AIDS Program
- Institute for BioNanotechnology in Medicine (IBNAM)
- Institute for Healthcare Studies

Feinberg School Contacts
- Search Faculty Profiles
- Administrative
- Academic
- Centers, Institutes, Programs and Affiliates
- Northwestern University
  - Directory Services
Departments, Centers & Institutes

Departments

- Anesthesiology
- Cell and Molecular Biology
- Dermatology
- Emergency Medicine
- Family and Community Medicine
- Medical Social Sciences
- Medicine
- Microbiology-Immunology
- Molecular Pharmacology and Biological Chemistry
- Neurological Surgery
- Neurology, Ken and Ruth Davee department
- Obstetrics and Gynecology
- Ophthalmology
- Orthopaedic Surgery
- Otolaryngology—Head and Neck Surgery
- Pathology
- Pediatrics
- Physical Medicine and Rehabilitation

Centers, Institutes, and Programs

- Bluhm Cardiovascular Institute
- Bushler Center on Aging, Health & Society
- Children’s Memorial Research Center

Feinberg School Contacts

- Search Faculty Profiles
- Administrative
- Academic
- Centers, Institutes, Programs and Affiliates
- Northwestern University
  Directory Services
Digitizing Folk Music History

Course Info

Schedule:
- Weeks 1 and 2
- Weeks 3 and 4
- Weeks 5 and 6
- Weeks 7 and 8
- Weeks 9 and 10
- Final

Syllabus (pdf file, 5/12/11)
An overview of our quarter.
Resource Not In Archive

The Resource you requested is not in this archive.
PRINCIPLE:
It is fair use to create topically based collections of websites and other material from the Internet and to make them available for scholarly use.

LIMITATIONS:

• Captured material should be represented as it was captured, with appropriate information on mode of harvesting and date.

• To the extent reasonably possible, the legal proprietors of the sites in question should be identified according to the prevailing conventions of attribution.

• Libraries should provide copyright owners with a simple tool for registering objections to making items from such a collection available online, and respond to such objections promptly.
Twitter: API available for just about all of it. Does not expressly prohibit crawlers (although robots.txt does)

Facebook: “You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission.” (API available).
SOCIAL MEDIA archiving and the story of the Twitter archive at the Library of Congress (with images of what social media may look like).
The Library of Congress Agreement with Twitter

The Library’s agreement with Twitter announced April 14, 2010 provided that:

• Twitter would donate a collection consisting of all public tweets from the Twitter service from its inception to the date of the agreement, an archive of 21 billion tweets that occurred between 2006 and 2010.

• Any additional materials Twitter provides to the Library would be governed by the terms of the agreement unless both parties agree to different terms in advance of receiving such additional materials.

• The Library could make available any portion of the collection six months after it was originally posted on Twitter to “bona fide” researchers.

• A researcher must sign a “notification” prohibiting commercial use and redistribution of the collection.

• The Library cannot provide a substantial portion of the collection on its web site in a form that can be easily downloaded.
Transfer of Data to the Library of Congress

• In December, 2010, Twitter named a Colorado-based company, Gnip, as the delivery agent for moving data to the Library.

• Shortly thereafter, the Library and Gnip began to agree on specifications and processes for the transfer of files – “current” tweets - on an ongoing basis.

• In February 2011, transfer of “current” tweets was initiated and began with tweets from December 2010.

• On February 28, 2012, the Library received the 2006-2010 archive through Gnip in three compressed files totaling 2.3 terabytes. When uncompressed the files total 20 terabytes. The files contained approximately 21 billion tweets, each with more than 50 accompanying metadata fields, such as place and description.

• As of December 1, 2012, the Library has received more than 150 billion additional tweets and corresponding metadata, for a total including the 2006-2010 archive of approximately 170 billion tweets totaling 133.2 terabytes for two compressed copies.
Behind the scenes—how to get your data out!
Twitter archive download

Your Twitter archive
Keep in mind that this download may contain sensitive content, so use caution before sharing it.

Your Twitter archive  Download

Your archive will be downloaded as a .zip file. Unzip the file and open 'index.html'. You may also receive an additional confirmation message from your web browser. If you do, click 'open' to proceed.
Twitter archive download

Jun 2014 3 TWEETS

Benn Joseph @doppelen
@bertramlyons @UWMadisonSLIS this is fantastic--thanks for sharing!
View on Twitter

Sarah Dorpinghaus @SMDorpinghaus
"Love doesn't pay the bills" D. Gracy on advocacy and archives #ALI14
tweeted by @doppelen
View on Twitter

Benn Joseph @doppelen
I just earned the 'I Believe in IPA! (Level 7)' badge on @untappd! untpd.it/s/b19615179 #ibelieveinIPA
View on Twitter

This is an offline archive of your Tweets from Twitter. Use the months above to navigate the archive.
Facebook archive download

Download Your Information
Get a copy of what you've shared on Facebook.

This is a copy of personal information you've shared on Facebook. To protect your info, we'll ask you to re-enter your password to confirm that this is your account.

Caution: Protect your archive
Your Facebook archive includes sensitive info like your private Wall posts, photos and profile information. Please keep this in mind before storing or sending your archive.

Download Archive
Benn Joseph

Profile
http://www.facebook.com/bennjoseph

Email
bennjoseph@gmail.com

Registration Date
Sunday, June 29, 2008 at 10:56am CDT

Birthday
08/25/1978

Gender
Male

Current City
Chicago, Illinois

Hometown
Warrenville, Illinois

Relationship Status
Married to Nath Wheaton since September 21, 2007

Family
Patrick Wheaton (Family member)
Robert Nicholas Wheaton (Uncle)
Lee Willie (Uncle)
Axelle Burot (Cousin (Pending))

Education
Wheaton Warrenville South HS 1996

Illinois State University 2000
Various arts

The University of North Carolina at Chapel Hill
MSLS
Information/Library Science

Employers
Northwestern University June 2009 to present
Manuscript Librarian

Chicago Historical Society March 2008 to June 2009
Benedictine University Library August 2006 to March 2008
Special Collections Librarian
Important information about your Google data archives

- Do not download your archives on public computers.
- It's important that you have control over your data. If you have the time, please let us know how we can improve.
- If you have decided to take your data elsewhere, please research the data export policies of your destination. Otherwise, if you ever want to leave the service, you may have to leave important stuff like your photos behind.

Archive of 16 Google products · 5.08 GB

Download this archive to your computer
Download each file to your computer. Keep these files safe since they contain your personal information.

Download 1/2  bennjoseph@...-takeout.zip · 924.1 MB

Download 2/2  All mail Including Spam and Trash-2.mbox · 4.18 GB
Download Google data

- What’s included?
“Save Page As...”
Technology and tools—how do we do it?
Tools

• “the Library of Congress is documenting ALL of Twitter. I work with SF/F authors and encourage them to either do periodic screencaps, printing to PDF, or imaging of some sort of their websites and blogs using what capture tools I can find. I also encourage them to submit their URLs to The Internet Archive in the hopes of additional snapshots.” –Lynne Thomas, Northern Illinois University.

• Wayback machine works for many websites, but not for social media.
Page cannot be crawled or displayed due to robots.txt.

Facebook robots.txt

# Notice: Crawling Facebook is prohibited unless you have express written permission. See: http://www.facebook.com/apps/site_scraping_tos_terms.php

User-agent: baiduspider
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /l.php
Disallow: /p.php
Disallow: /photo.php
Disallow: /photo_comments.php
Disallow: /photo_search.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: Googlebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /l.php
Disallow: /p.php
Disallow: /photo.php
Disallow: /photo_comments.php
Disallow: /photo_search.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: ia_archiver
Allow: /about/privacy
Allow: /full_data_use_policy
Allow: /legal/terms
Allow: /policy.php
Disallow: /

User-agent: *
Disallow: /
Tools

• All-inclusive tools commonly used by academic institutions: WAS and Archive-It.
• WAS works good for simple websites, but not social media.
• Archive-It might work (Umbra).
• Neither allow for much customization of crawl options. What else is there?
Command-line interfaces (wget, curl, and others) are more customizable:

- **(STEP1)**
  
  ```
  mkdir /websitedl/
  cd /websitedl/
  ```

- **(STEP2)**
  
  ```
  ```
Facebook and Twitter both have APIs (use cURL)

Facebook’s Graph API named after the idea of a social graph—a representation of the information on Facebook composed of:

- **nodes** (basically "things" such as a User, a Photo, a Page, a Comment)
- **edges** (the connections between those "things", such as a Page's Photos, or a Photo's Comments)
- **fields** (info about those "things", such as the birthday of a User, or the name of a Page).
Tools

• **HTTrack** (Windows—allows for similar commands as wget)
• **Sitecrawler** (Mac)
• **Social Feed Manager** (developed at George Washington University, a django application for managing multiple feeds of social media data)
Experimental tools

WARCreate, a Chrome plugin

WAIL (Web Archiving Integration Layer)

Archive Facebook (Firefox plugin)
Tools: subscription services

Using subscription services can be much easier—but also much more expensive!
So...there are many challenges.


• What methods will you use for collecting your data?
• How often will you collect it?
• Where will you put it?
• How will you provide access to it?
Sustaining the collection

Guide to the Northwestern University Web Archive

Collection Title: Northwestern University Web Archive
Dates: 2001-2014
Identification: 77/1
Creator: Northwestern University (Evanston, Ill.)
Language of Materials: English
Abstract: The Northwestern University Web Archives is an initiative managed by the Northwestern University Archives which crawls and makes available websites and other online content which falls within the institutional collecting scope.

Acquisition Information: Sites in the Northwestern University Web Archive have been acquired with the California Digital Library’s Web Archiving Service, which uses the Heritrix web crawler.

Conditions Governing Access: No restrictions.
Repository: Northwestern University Archives
Deering Library, Room 110
1970 Campus Dr.
Evanston, IL, 60208-2300
URL: http://www.library.northwestern.edu/archives
Email: archives@northwestern.edu
Phone: 847-491-3354

Scope and Content
The Northwestern University web archives are intended to document and preserve online resources of unique, essential, and enduring value. These archived websites parallel and complement NUA’s manuscript collections and record groups (both paper-based and digital). Many important documents are now available exclusively online and the expansion of interactive features, forms, and collaborative resources have made websites indispensable to the university’s daily operation. Although certain content types and features are difficult to capture and render accurately, NUA aims to preserve academic and administrative policies; significant publications and resources; important representations of research, instruction, and creative work; and the overall look and feel of the Northwestern University web domain.
<table>
<thead>
<tr>
<th>Administration</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration and Planning Web Archives</td>
<td>Item 1</td>
</tr>
<tr>
<td>Title: Administration and Planning Web Archives</td>
<td></td>
</tr>
<tr>
<td>African American Student Affairs Web Archives</td>
<td>Item 2</td>
</tr>
<tr>
<td>Title: African American Student Affairs Web Archives</td>
<td></td>
</tr>
<tr>
<td>Center for Student Involvement Web Archives</td>
<td>Item 3</td>
</tr>
<tr>
<td>Title: Center for Student Involvement Web Archives</td>
<td></td>
</tr>
<tr>
<td>Commencement Web Archives</td>
<td>Item 4</td>
</tr>
<tr>
<td>Title: Commencement Web Archives</td>
<td></td>
</tr>
<tr>
<td>Counseling and Psychological Services (CAPS) Web Archives</td>
<td>Item 5</td>
</tr>
<tr>
<td>Title: Counseling and Psychological Services (CAPS) Web Archive</td>
<td></td>
</tr>
<tr>
<td>Discover: Research at Northwestern Web Archive</td>
<td>Item 6</td>
</tr>
<tr>
<td>Title: Discover: Research at Northwestern Web Archive</td>
<td></td>
</tr>
<tr>
<td>Division of Student Affairs Web Archives</td>
<td>Item 7</td>
</tr>
<tr>
<td>Title: Division of Student Affairs Web Archives</td>
<td></td>
</tr>
<tr>
<td>Fraternity and Sorority Life Web Archives</td>
<td>Item 8</td>
</tr>
<tr>
<td>Title: Fraternity and Sorority Life Web Archives</td>
<td></td>
</tr>
<tr>
<td>Graduate Housing Web Archives</td>
<td>Item 9</td>
</tr>
<tr>
<td>Title: Graduate Housing Web Archives</td>
<td></td>
</tr>
<tr>
<td>Living Off-Campus Web Archives</td>
<td>Item 10</td>
</tr>
<tr>
<td>Title: Living Off-Campus Web Archives</td>
<td></td>
</tr>
<tr>
<td>Living@NU Web Archives</td>
<td>Item 11</td>
</tr>
<tr>
<td>Title: Living@NU Web Archives</td>
<td></td>
</tr>
<tr>
<td>Multicultural Student Affairs Web Archives</td>
<td>Item 12</td>
</tr>
<tr>
<td>Title: Multicultural Student Affairs Web Archives</td>
<td></td>
</tr>
</tbody>
</table>
Sustaining the collection

- Bagit specification
- Using checksums
- Creating durable digital objects
Sustaining the collection

• How would you talk about the costs of a web archiving or social media archiving project?
Thank you!

Benn Joseph
b-joseph@northwestern.edu
How does the federal government collect social media?

Methods to capture social media records include:

- Using web crawling or other software to create local versions of sites;
- Using web capture tools to capture social media;
- Using platform specific application programming interfaces (APIs) to pull content;
- Using RSS Feeds, aggregators, or manual methods to capture content; and
- Using tools built into some social media platforms to export content.
LC NDSA also gives “personal archiving” guidance for websites, blogs, and social media.

- **Identify** where you have Websites
- **Decide** which information has long-term value
- **Export** the selected information
- **Organize** the information
- **Make** copies and *manage* them in different places