



Archiving the Social Web

MARAC Spring 2013 Conference

April 2013

Lori Donovan
Partner Specialist
Internet Archive



About Internet Archive

We are a Digital Library

Mission Statement: Universal access to all knowledge

- Founded by Brewster Kahle in San Francisco, California in 1996
- Officially designated a Library by the State of California in 2007



Access to General Web Archive

The Archive is accessible to the public via the website: www.archive.org

- Started collecting content in 1996
- 280+ billion web pages
- 80+ million websites
- Almost every domain
- Content in 40+ Languages
- Aggregate a broad snapshot of the web every two months - approximately 3 billion pages a snapshot



About Archive-It

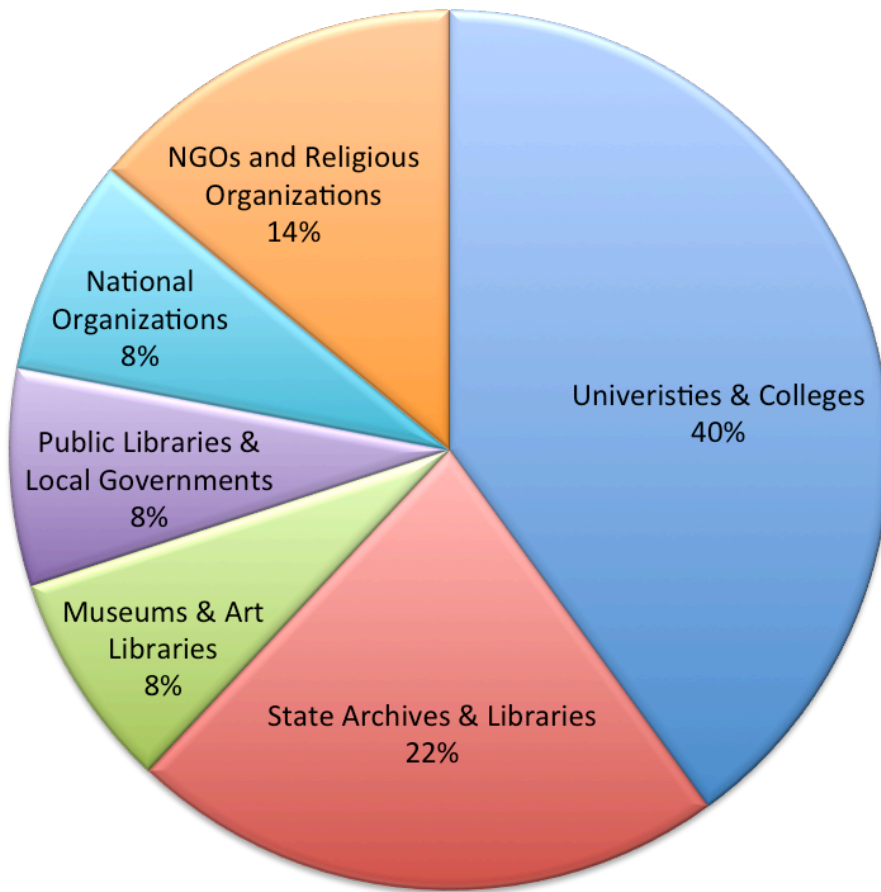
Archive-It is a subscription service deployed in February 2006

- **Web based application** that allows users to create, manage, access and store collections of digital content
- The service is a **fully hosted solution**, and includes access and storage.
- **Provides tools for selection and scoping** including cataloging with metadata
- Ability to **capture content using 10 different crawl frequencies**
- Archived content includes: html, videos, audio, PDF, images, social media, online newspapers
- **Can browse archived content 24 hours after a capture is complete**; and full text search is available within 7 days
- **Restricted access options** are available



Archive-It Partners

247 Partners in 46 U.S. States and 16 Countries



- Universities & Colleges
- State Archives & Libraries
- Museums & Art Libraries
- Public Libraries & Local Governments
- National Organizations
- NGOs and Religious Organizations



Why our Partners are archiving Web Content

- The web and specifically social networking sites have transformed how people communicate and receive information
- The availability of electronic information is taken for granted and it is a fallacy that if information is on the web it will be there forever.
- The web represents who we are. It's our culture and an emerging means of social engagement, and we don't want to lose it.



Social Media Overview

- Capturing social media sites is becoming more necessary for organizations archiving the web
- Focused on: Flickr, Facebook, Twitter, and YouTube
- On our radar: Vimeo, LinkedIn, Google+



Why Archive Social Media Sites?

- **State Agencies:** An increasing number have decided that the content on these sites are a record and need to be archived. "A tweet is a record"
- **University libraries:** Used to share information with students and alumni and contain important records about a school's culture, student body and campus events.
- **Non Government Non Profit Organizations:** Used to record online presence and impact
- **Researchers:** Used to preserve valuable social reactions and change on topics of interest



North Carolina State Archives & State Library of North Carolina

Purpose: archive state agency websites and publications

- Includes pages in a variety of formats: text, images, audio, video and social networking sites
- Were among the first Archive-It partners to archive social networking sites



North Carolina State Archives & State Library of North Carolina

facebook

Keep me logged in

[Forgot your password?](#)

Email

Password

Login

Sign Up

Bev Perdue is on Facebook

Sign up for Facebook to connect with Bev Perdue.



Bev Perdue

Wall Info Photos Notes

Bev Perdue + Fans **Bev Perdue** Just Fans



Bev Perdue

A celebration of struggle and continued achievement – Blog by Jill Dinwiddie Ex. Dir., NC Council for Women/Domestic Violen...

The Council for Women is proud to have been a partner in a Women's History Month Celebration that truly honored the pioneering women who led the suffragist movement, and whose victories paved the way for today's women to earn their rightful place in a wide range of leadership roles. I...

43 minutes ago · Comment · Like · View Feedback (8)



Bev Perdue

Building a better future for North Carolina – Blog by Moses Carey Secretary of Administration

Earlier today I enjoyed the opportunity to welcome nearly 1,200 participants of the 29th annual State Construction Conference at the McKimmon Center. Through the State Construction Office, we provide a forum for industry leaders – and hopefuls – to gather each year to brainstorm idea...

43 minutes ago · Comment · Like · View Feedback (8)



Bev Perdue

Gov. Perdue Announces Some North Carolina Farmers Eligible for Federal Disaster Assistance – Press Release

Gov. Perdue today announced that U.S. Secretary of Agriculture Tom Vilsack has designated much of Eastern North Carolina as a disaster area after excessive rain and flooding in recent months caused significant crop losses...

Posts and comments to and from this site, in connection with the transaction of public business, are subject to the North Carolina Public Records Law and may be disclosed to third parties.

Information

Current Office

Office:

Governor

State:

North Carolina



Library of Virginia

Purpose: Preserve websites relating to Virginia government and elections

- Collection on current Governor includes Twitter and Flickr sites
- Collection on Twitter, Flickr, and Facebook sites of politicians and political organizations in Virginia



Governor of Virginia's photostream pro

[Photostream](#) | [Sets](#) | [Favorites](#) | [Galleries](#) | [Profile](#) | [More](#) ▾

Slideshow Share ▾



IMG_2137

© All rights reserved
Uploaded on [Apr 16, 2013](#)
15 views



IMG_2134

© All rights reserved
Uploaded on [Apr 16, 2013](#)
11 views

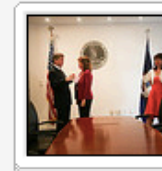


IMG_2129



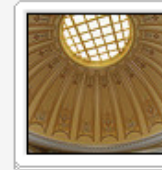
[Y:\Media Resource...](#)

28 photos
267 views



[Governor Meets With...](#)

24 photos
162 views



[Veto Session - April 3, 2013](#)

21 photos
215 views



[HB 2262 Bill Signing...](#)

5 photos
55 views



Have an account? [Sign in](#)

Get short, timely messages from Governor of Virginia.

Twitter is a rich source of instantly updated information. It's easy to stay updated on an incredibly wide variety of topics. [Join today](#) and follow [@governorva](#).

[Join Today >](#)

Get updates via SMS by texting [follow governorva](#) to 40404 in the United States
[Codes for other countries](#)



governorva

Virginia Secretary of Health and Human Resources Dr. Bill Hazel Announces Virginia Health Reform Initiative... <http://bit.ly/aENDIU>

2:29 PM May 14th via Facebook

For First Time in Two Years, Virginia Posts Consecutive Months of Revenue Growth... <http://bit.ly/ak1D3d>

9:22 AM May 13th via Facebook

Governor McDonnell Announces \$492.7 Million Transportation Bond Sale... <http://bit.ly/chACjv>

12:44 PM May 12th via Facebook

Virginia Film Office Wins Major Marketing Awards at Los Angeles Trade Event... <http://bit.ly/bPJ1gs>

8:21 AM May 12th via Facebook

Governor McDonnell Announces Significant Investment in Chesapeake Bay Clean-Up... <http://bit.ly/cqup3N>

1:10 PM May 11th via Facebook

Governor McDonnell Unveils Biotech Re-Entry Initiative

Name Governor of Virginia

0 following 478 followers 61 listed

Tweets 74

Favorites

Following

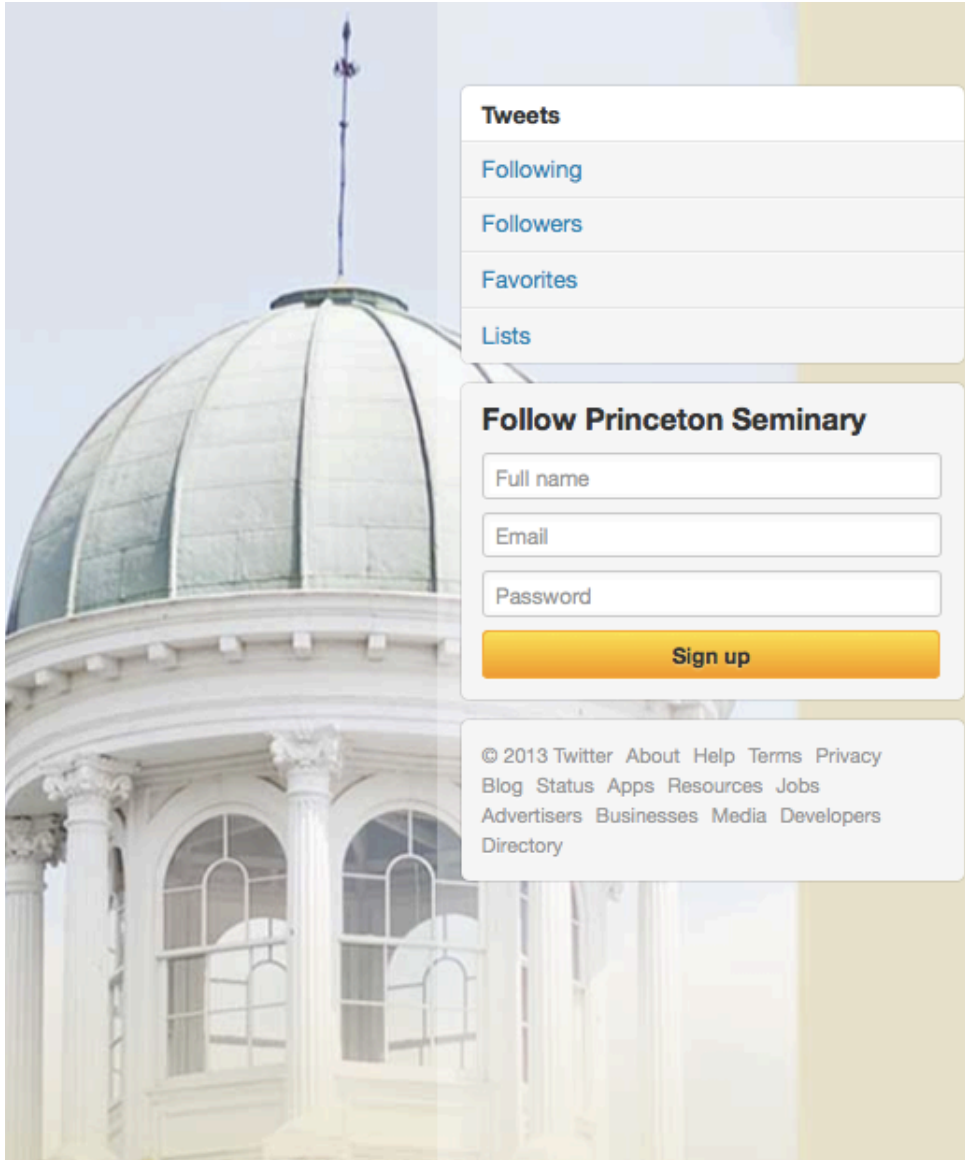
[RSS feed of governorva's tweets](#)



Princeton Theological Seminary

Purpose: Archive content created by Princeton Theological Seminary as part of its official institutional identity on the web.

- Archiving Facebook, Twitter, YouTube
- Partner since 2012



- Tweets**
- [Following](#)
- [Followers](#)
- [Favorites](#)
- [Lists](#)




Follow Princeton Seminary

© 2013 Twitter [About](#) [Help](#) [Terms](#) [Privacy](#)
[Blog](#) [Status](#) [Apps](#) [Resources](#) [Jobs](#)
[Advertisers](#) [Businesses](#) [Media](#) [Developers](#)
[Directory](#)



Princeton Seminary
@ptseminary
Princeton Seminary prepares women and men to serve Jesus Christ in ministries by equipping them for leadership worldwide.
Princeton, New Jersey · <http://www.ptsem.edu>

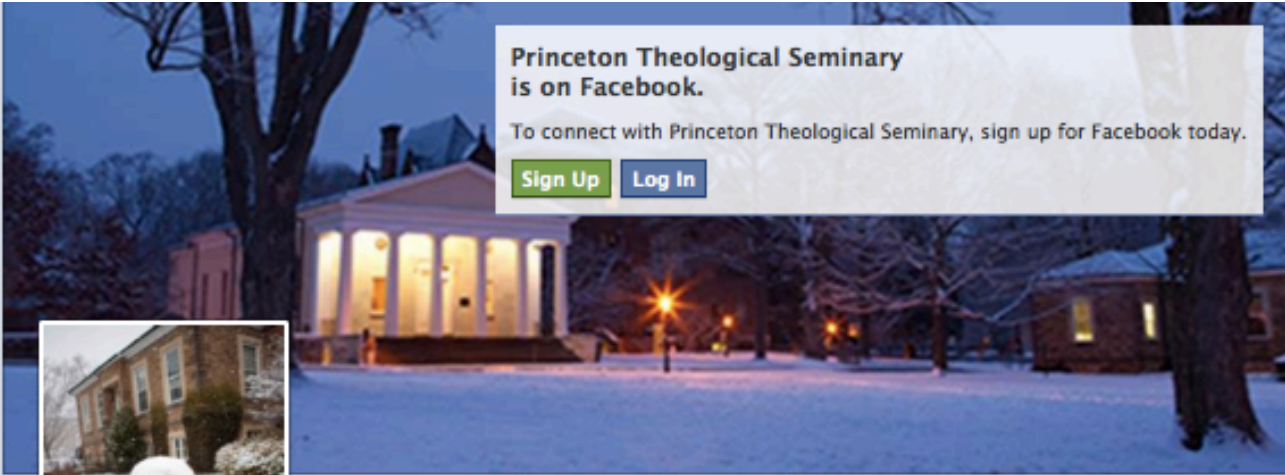
261 TWEETS	29 FOLLOWING	382 FOLLOWERS	<input type="button" value="Follow"/>
----------------------	------------------------	-------------------------	---------------------------------------

- Tweets**
-  **Princeton Seminary** @ptseminary 4 Feb
Join us for the 2013 Princeton Alumni/ae Conferences! ptsem.edu/princetonconfe... #ptseminary
[Expand](#)
 -  **Princeton Seminary** @ptseminary 1 Feb
Join us for special events during Black History Month! ptsem.edu/index.aspx?id=...
[Expand](#)
 -  **Princeton Seminary** @ptseminary 31 Jan
Alumni/ae transcripts can now be ordered online! our.ptsem.edu/ics/Campus_Lif...
[Expand](#)

Princeton Theological Seminary is on Facebook.

To connect with Princeton Theological Seminary, sign up for Facebook today.

Sign Up Log In



Princeton Theological Seminary

1,895 likes · 59 talking about this · 3,141 were here

Like

College & University The official Facebook page of Princeton Theological Seminary.



1,895



About

Photos

Likes

Events

Map

Highlights

Princeton Theological Seminary created an event. 2 hours ago. Lutherans in Diaspora Conference Keynote Speech: "I was in Prison": Re-membering Jesus in Lockdown America. February 8 at 6:45pm in EST. Lutheran Church of the Messiah. Princeton Theological Seminary is going. Like · Comment · Follow Post · Share

Princeton Theological Seminary created an event. 2 hours ago. Black History Month Event: "Emancipation of the Bodies: The Mass Incarceration Film Documentary Screening and Panel Discussion"

Recommendations See All. Daniel Philip Calder My name is Daniel Calder, and I am a student of over ... 2 · about 2 weeks ago. Tom Stephen Erin Dunigan a evangelist for the Presbyterian Church... about 2 months ago. Jason Bruner If I may make another recommendation. I recommend... 1 · about 5 months ago. Jason Bruner In light of the theft of one SUV, a bicycle, and 2 bicycl... 2 · about 5 months ago. See More



Twitter

What types of pages can be archived?

– Individual user feeds

- <https://twitter.com/archiveitorg/>

– Searches

- <https://twitter.com/search?q=web%20archiving&src=typd>

– Lists

- <https://twitter.com/smithsonian/smithsonian/>

– A specific tweet

- <https://twitter.com/archiveitorg/status/294819565320413184>



Links in Tweets

- Can I archive a URL linked to using a ‘URL shortener’ ?
 - Yes! Use an Expand Scope rule for <http://t.co/> - all URLs posted on Twitter redirect through that domain
 - Note: just the one page that the URL shortener link points to will be archived (plus embedded content)



Facebook

What types of pages can be archived?

- Individual User Profiles – Timeline view
 - <http://www.facebook.com/tonyforsenate/>
- Pages - Timeline view
 - <http://www.facebook.com/ArchiveIt/>
- Events
 - <http://www.facebook.com/events/265897963430841/>
- Albums
 - <https://www.facebook.com/media/set/?set=a.13499334573.18616.6193904573&type=3>



Facebook

What about dynamically loading content?

- Currently we can capture the initial content on a Facebook timeline, however the dynamically loading content can be difficult to capture due to the way that content is served by Facebook
- Our engineers are working on keeping up to date with these changes and we are also investigating alternate methods for capturing Facebook pages



YouTube

What types of pages can be archived?

- Channel /User pages
 - <http://www.youtube.com/whitehouse>
- Watch pages- individual videos
 - http://www.youtube.com/watch?v=5IVluW8vJ_E
- Uploaded Document RSS Feed
 - <http://gdata.youtube.com/feeds/api/users/whitehouse/uploads/>
- Embedded YouTube Videos on other sites:
 - <http://www.whitehouse.gov/photos-and-video/video/2013/01/29/president-obama-speaks-comprehensive-immigration-reform>



Flickr

What types of pages can be archived?

- Photo streams

- Ex: <http://www.flickr.com/photos/whitehouse/>

- Individual photos

- Ex: <http://www.flickr.com/photos/whitehouse/8390033709/in/photostream>



Other Sites

- Can sites other than those already mentioned be archived?
 - Yes! There are many more sites out there that can be archived.
 - Other sites mentioned by partners currently are Google+, LinkedIn, Vimeo, and SlideShare.



Typical Challenges

- Most crawlers cannot capture content behind log-ins currently
 - Feature in Archive-It 4.8 Release, April 2013
- Some parts of sites are not “archive-friendly” (i.e. complex javascript, Flash, etc.)
- Social Media sites tend to change both their technical structure and policy quickly and often.
- Structure of the sites/URLs means users need to add scoping rules to only capture content you are interested in. Each site has its own unique set of challenges.



Overall Approaches

- Trial and Error: Try to harvest with a variety of settings and a variety of seeds
- Quality Review: review archived content thoroughly
- Collaborate: compare approaches and results with other Archive-It users
- Document detailed instructions, lessons learned, and best practices for other partners



Moving Forward

- These best practices will change as the sites themselves make changes. We continually update Help pages as settings change
- We continue to focus on working with our partners to improve the capture and display of archived social networking sites
- The Archive-It team is exploring other capture mechanisms besides using a traditional crawler resource (Heritrix)
 - Headless browsers
 - Hybrid architecture
 - API
 - Partnering with third party software
- Enhance the display and search capabilities