Archiving Social Media for State Agencies

MARAC
Michael P. Martin
New York State Archives
April 26, 2013
The First Crawl 2006-2007

- Comprehensive approach
  - Capture all public State government sites
- Choose crawler
  - Went with OCLC Web Harvester
- Had to identify the sites
  - Over 300 official NY State agency websites
- Had to set up workflow
Workflow

- Set up crawl
  - Site URL
  - Crawl parameters
    - Depth level of 10
  - Agency’s name
  - Site’s title
- Review results
- Transfer results to OCLC Digital Archive
- Follow up with OCLC if needed
First Time Jitters: 2006-2007 Crawl

- Working for the first time
- Working with Beta version of WAW
  - Missed some larger sites
- Did not even think about social media sites
2008 Crawl
Problems that Kept Reoccurring

- Crawlers can’t capture everything
  - Content accessed by completing a form
  - Some types of Javascript
  - External stylesheets and images
  - Most multimedia content
- Must be in office to start a crawl
Pilot Spring 2010

- **Governor**
  - Twitter
  - Facebook
  - Flickr
  - Youtube

- **New York State Archives**
  - Facebook
  - Twitter

- **Results**
  - robots.txt exclusion override
  - Not much more than top page
  - Otherwise it would try to capture all of social media site
Identifying Social Media Sites

- **Method**
  - Start with results of previous crawls
  - Examine each main site for links
  - Google searches

- **Results**
  - Identified dozens of sites
  - Created our own database
## OCLC Web Harvester

### Harvest Parameters Set Up
- **URL:** https://www.facebook.com/GovernorAndrewCuomo
- **Ignore Robots:** Yes
- **Harvest Type:** By Links
- **Harvest Depth:** 0
- **Preview Harvest:** No
- **Select a CONTENTdm collection for this web content:** 2012-2013 Archived Web Sites

### Serial Information
- **Are you harvesting an item in a serial?** No
- **Issue Title**
- **Issue Date (YYYY-MM-DD)**

### Notification
- **Email Address:** mmartin@mail.nysed.gov

(If you do not select the email notification option, you can check the status of your harvests using the Review Web Harvests menu option.)
Archived New York State Government Web sites

2012 Archived Web Sites
Archival copies produced in 2012 as a result of agency mergers or creation of new sites.

2010-2011 Archived Web Sites
Archival copies produced as Governor David A. Paterson (2008-2010) left office. Most of these copies were created in December 2010 and January 2011, but a few were completed in the months that followed.

2010 Archived Web Sites
Archival copies produced in May and June 2010 as a result of agency mergers or the creation of new sites.

About the collections
This site provides access to archival copies of New York State government Web sites produced from December 2006 onward.
CONTENTdm Search Screen

- **Description**: Harvested from the web on 3/12/10 (State Archives accession B1972-10)
- **Creator**: New York (State). Governor (2008-2010: David A. Paterson)
- **Source**: http://twitter.com/MPaigePaterson
New York State Archives

The New York State Archives leads efforts to manage, preserve, ensure open access to, and promote the wisdom of records that support information needs and document the history, governments, events and peoples of our State.

Information
- Locations
  - Cultural Education Center
    - Albany, NY, 12200

New York State Archives

New York State Archives is on Facebook
Sign up for Facebook to connect with New York State Archives.

New York State Archives

New York State Archives
A new slide show, "Winter in New York" is now available on the Archives web site. Check it out at http://www.archives.mnhs.org/slide.html

New York State Archives Homepage
www.archives.mnhs.org
December 16, 2009 at 6:47am

Donna, Arriet, Arthur and 6 others like this.

Prudence Beckman
Some of these images are amazing! It's great to have them see the light of day.
December 16, 2009 at 6:38am

Meredith Clever Advertising:
I have one of the images saved at my desktop! The Beacon is a treasure - thanks for keeping them coming!
December 15, 2009 at 9:17pm

New York State Archives

New workshop Electronic Data Storage
Come to this non-technical workshop intended to provide records managers with a basic overview of common ways electronic data is stored and an understanding of how it fits in with records management. This workshop will be held in Albany on December 15, 2009.
December 8, 2009 at 10:16am
More problems

- Changes in sites caused problems
- Could not capture what we had before
Youtube
Flickr
Get short, timely messages from New York Senate.

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 @NYSenate.

NYSenate

RT @NYSenateCIO: Announcing @NYSenate's OpenLegislation 1.6 "ember eltanin" http://bit.ly/h0HGQz #OpenNY #OpenGov #Gov20
11:04 AM Dec 22 via TweetDeck

2nd meeting of Senate Leg Task Force on Demographic Research & Reapportionment under way. Live vid: http://bit.ly/fIWoLq #redistricting
9:06 AM Dec 16th via TweetDeck

8:41 AM Dec 16th via TweetDeck

Ideascale Blogs

Straight Talk from the Taxpayer: Scheduling Online Payments

On December 27, 2010, In Straight Talk

Taxpayer Post: "I suggest that Online Services allow a taxpayer to schedule a date to pay taxes. This would make paying sales and income taxes easier."

Response: Currently, taxpayers can schedule a number of on-line payments in advance through the New York State Tax Department’s secure web site. These include estimated taxes and extensions (for both personal income tax and corporation tax) and payments of the

Thank You for Visiting Straight Talk, Straight Answers
This blog provides information, feedback and resources in response to comments submitted by taxpayers. This open and transparent dialogue is important to Governor Paterson’s ongoing efforts to save taxpayer dollars, streamline government and provide relief to New York’s property taxpayers. Please visit our companion site Straight Talk from the Taxpayer to submit, discuss and vote on ideas.
Other Tools to Capture Content

- **Facebook**
  - Allows you to download your own info
  - Can be quite extensive

- **ArchiveFacebook**
  - Firefox Add on

- **ArchiveSocial**
Now and Next Steps

- Making results accessible
  - Catalog records
  - Noting problems

- Plan for future crawls
  - Yearly for some sites
  - Keep testing
  - 2014
    - Next major crawl
Contact

Michael P. Martin
Electronic Records Archivist
New York State Archives
9D64 Cultural Education Center
Albany, NY 12230
518-486-1741
mmartin@mail.nysed.gov