

---

# Archiving Social Media for State Agencies



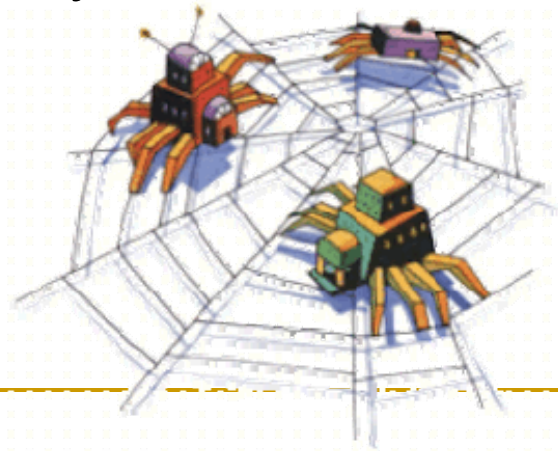
---

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

The screenshot shows the 'Web Harvester' interface with the 'Setup' tab selected. The interface includes a title field (currently empty), an OCLC Number field, and a 'Harvest Parameters Set Up' section with the following options:

- URL: [Text input field]
- Ignore Robots?: Yes (selected)
- Harvest Type: By Links (selected)
- Harvest Depth: 0 (selected)
- Preview Harvest?: No (selected)
- Select a CONTENTdm collection for this web content: 2012-2013 Archived Web Sites (selected)

The 'Serial Information' section includes:

- Are you harvesting an item in a serial?: No (selected)
- Issue Title: [Text input field]
- Issue Date (YYYY-MM-DD): [Text input field]


The 'Notification' section includes:

- Email Address: mmartin@mail.nysed.gov

At the bottom, there are buttons for HARVEST, CLOSE, CLEAR, and HELP.

# 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

BETA  , 100314535

Discovery Tool Properties Tool Analysis Tool Harvest Tool

Status Quick Harvest

Quick Harvest Setup

What do you want to harvest?

Harvest

Title \*: New York State Archives [website]

Website \*: <http://www.archives.nysed.gov/aindex.shtml>

Entity: New York State Archives.

OCLC Number:

Spider Settings: NYS Web Crawl Harvest

Elements marked with

Harvest

- <Default>
- NYS Harvest Top Page Only
- NYS Web Crawl Harvest
- Test NYS Web Crawl Harvest



# 2008 Crawl



---

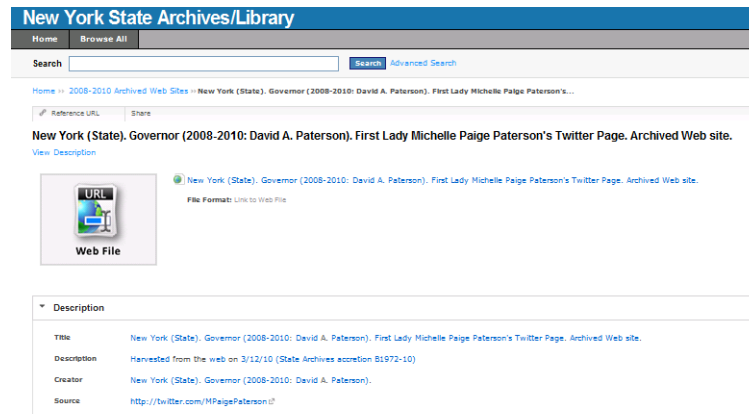
# Problems that Kept Reoccurring

- 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



The screenshot shows the New York State Archives/Library website. At the top, there is a navigation bar with 'Home' and 'Browse All'. Below that is a search bar with a 'Search' button and a link to 'Advanced Search'. The main content area displays a search result for 'New York (State), Governor (2008-2010: David A. Paterson), First Lady Michelle Paige Paterson's Twitter Page. Archived Web site.' The result includes a 'Web File' icon, a description, and a table with the following information:

Description	
Title	New York (State), Governor (2008-2010: David A. Paterson), First Lady Michelle Paige Paterson's Twitter Page. Archived Web site.
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	<a href="http://twitter.com/MPaigePaterson/">http://twitter.com/MPaigePaterson/</a>



The screenshot shows the Facebook page for the New York State Archives. The page header includes the Facebook logo, a search bar, and a 'Log Out' button. The main content area features a profile picture of the Archives, a cover photo, and a 'Wall' section with several posts. The posts include text updates and photos, such as one about a new 1850 show 'Victor in New York' and another about a workshop on electronic data storage.

---

# 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

Most Visited e-Library OPAC iLink at... LATS Login Archives Manag... CONTENTdm Collectio... Google NY:Gov Login Social Media Archiving... Digitization 101

## Web Harvester

Setup

Review

Reports

Title: No Title Provided

OCLC Number:

### Harvest Parameters Set Up

URL	<input type="text" value="https://www.facebook.com/GovernorAndrewCuomo"/>
Ignore Robots?	<input type="button" value="Yes"/>
Harvest Type	<input type="button" value="By Links"/>
Harvest Depth	<input type="button" value="0"/>
Preview Harvest?	<input type="button" value="No"/>
Select a CONTENTdm collection for this web content:	<input type="button" value="2012-2013 Archived Web Sites"/>

### Serial Information

Are you harvesting an item in a serial?	<input checked="" type="radio"/> No <input type="radio"/> Yes
Issue Title	<input type="text"/>
Issue Date (YYYY-MM-DD)	<input type="text"/>

### Notification

Email Address	If you would like to be notified about the status of your harvest, please enter an email address: <input type="text" value="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.)

# CONTENTdm Collections Page

## New York State Archives/Library

[Home](#) [Browse All](#)

Search

[Search](#)

[Advanced Search](#)

### Archived New York State Government Web sites

#### All Collections

##### [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

**New York State Archives/Library**

Home Browse All Log In Help

Search  within results [Search](#) [Advanced Search](#)

**Add or remove other collections to your search:**

- 2006-2010 Archived Web Sites
- 2008-2010 Archived Web Sites
- 2010 Archived Web Sites
- 2010-2011 Archived Web Sites
- 2012 Archived Web Sites
- 2012-2013 Archived Web Sites

[Hide](#)

**Narrow your search by:**






- Creator
- Description

**You've searched:** All Collections

All fields: twitter

Sort by: Description [Display Options](#) [Previous](#)

Display: 20

Thumbnail	Title	Description	Format	Creator
	New York Racing Association. Twitter (Andy Serling). Archived Web site.	Harvested from the web on 1/4/13 (State Archives accretion B2356-13)		New York Racing Association.
	New York (State). Legislature. Senate. Twitter. Archived Web site.	Harvested from the web on 1/7/13 (State Archives accretion B2197-13)		New York (State). Legislature. Senate.
	New York (State). Legislature. Senate. Twitter. Archived Web site.	Harvested from the web on 12/29/10 (State Archives accretion B2197-11)		New York (State). Legislature. Senate.
	New York (State). Governor's Office of Taxpayer Accountability. Twitter Page. Archived Web site.	Harvested from the web on 12/29/10 (State Archives accretion B2279-10)		New York (State). Governor's Office of Taxpayer Accountability.
	New York (State). Governor (2008-2010: David A. Paterson). First Lady Michelle Paige Paterson's Twitter. Archived Web site.	Harvested from the web on 12/31/10 (State Archives accretion B1972-10A)		New York (State). Governor (2008-2010: David A. Paterson).

# CONTENTdm Metadata

**New York State Archives/Library**

Home Browse All

Search  Search Advanced Search

Home >> 2008-2010 Archived Web Sites >> New York (State). Governor (2008-2010: David A. Paterson). First Lady Michelle Paige Paterson's...

Reference URL Share

**New York (State). Governor (2008-2010: David A. Paterson). First Lady Michelle Paige Paterson's Twitter Page. Archived Web site.**

[View Description](#)

 **Web File**

 [New York \(State\). Governor \(2008-2010: David A. Paterson\). First Lady Michelle Paige Paterson's Twitter Page. Archived Web site.](#)

**File Format:** Link to Web File

▼ **Description**

<b>Title</b>	<a href="#">New York (State). Governor (2008-2010: David A. Paterson). First Lady Michelle Paige Paterson's Twitter Page. Archived Web site.</a>
<b>Description</b>	Harvested from the web on 3/12/10 (State Archives accession B1972-10)
<b>Creator</b>	<a href="#">New York (State). Governor (2008-2010: David A. Paterson).</a>
<b>Source</b>	<a href="http://twitter.com/MPaigePaterson">http://twitter.com/MPaigePaterson</a>

# Captured Website - Facebook

**facebook**  Keep me logged in [Forgot your password?](#)

Email  Password

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



**NEW YORK**  
**archives**

**LINCOLN IN NEW YORK**  
WITH GREAT GUYTON HARRIS SLIDES

WILLIAM KENNEDY — JAMES W. BURNHAM — CRAIG L. STUBBS — CATHERINE CLAYTON  
— PEARL J. WILLIAMS — JAMES HILFER BOSTER — RAND L. THOMPSON

**New York State Archives**

Wall Info News YouTube Photos Boxes

**Just Fans**

**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.nysed.gov/aindex.shtml>

**New York State Archives Homepage**  
[www.archives.nysed.gov](http://www.archives.nysed.gov)  
December 16, 2009 at 6:47am

Donna, Albert, Arthur and 8 others like this.

**Prudence Backman** Some of these images are amazing! It's great to have them see the light of day.  
December 16, 2009 at 8:28am

**Meredith Cherven-Holland** I have one of the images saved as my desktop! The lantern slides are a visual treasure - thanks for keeping them coming!  
Mon at 8:17pm

**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 10:55am

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

**Information**

Location:  
Cultural Education Center  
Albany, NY, 12230

# Showtime Dec. 2010 - Jan. 2011



# More problems

- Changes in sites caused problems
- Could not capture what we had before

## Oops

Something went wrong. We're working on getting this fixed as soon as we can. You may be able to try again.

Okay

Whoops! Something went wrong. Please try again!

If you have a success story to share please send it to [mwbe@cio.ny.gov](mailto:mwbe@cio.ny.gov) #nysmwbe

8:28 AM Nov 16th via TweetDeck

You can find more information about the report at [www.nylovesmwbe.ny.gov](http://www.nylovesmwbe.ny.gov)

8:18 AM Nov 16th via TweetDeck

## Redirect Notice

This resource is outside the scope of the harvested document. Click below to access this resource on the Internet.

<http://www.flickr.com/photos/governordavidapateron/5257849465/>

Note: This resource may not be available on the Internet.

Click [back](#) to return to the harvested document.

# Youtube

The screenshot shows a YouTube channel page for 'nysoasas'. The channel name is 'nysoasas' with the subtitle 'nysoasas's Channel'. A yellow 'Subscribe' button is visible. The navigation menu includes 'All', 'Uploads', and 'Favorites'. The main video player area is currently black. Below the player are buttons for 'Info', 'Favorite', 'Share', 'Playlists', and 'Flag'. The 'Playlists' button is highlighted with a dashed border. To the right of the player is a search bar and a list of video thumbnails. The video list includes:

- Opening Celebration - Beyoncé (430 views - 11 months ago, 1:45)
- Opening Celebration - Beyoncé (533 views - 11 months ago, 1:21)
- Taking Action in NY: Introduction (225 views - 1 year ago, 1:27)
- Taking Action in NY: Part 1 - Strengthening (34 views - 1 year ago, 1:59)
- Taking Action in NY: Part 2 - Changing (40 views - 1 year ago, 2:46)
- Taking Action in NY: Part 3 - Creating (12 views - 1 year ago, 2:39)
- Taking Action in NY: Part 4 - Enforcement: A



# Flickr

Flickr: NYSLabor's Photostream - Windows Internet Explorer provided by NYS Education Department

http://worldcat.org/arcviewer/4/AO%23/2011/01/03/H1294085203555/viewer/file2.html

File Edit View Favorites Tools Help

Favorites Flickr: NYSLabor's Photostream

flickr from YAHOO!

Home The Tour Sign Up Explore Upload

You aren't signed in Sign in Help


Search NYSLabor's photostream Search

### NYSLabor's photostream

Sets Galleries Tags People Archives Favorites Profile

Slideshow


**CareerZone - Myers Middle School**



Smiles say it all - CareerZone is a fun way to learn!

All rights reserved  
Uploaded on Jun 23, 2010  
0 comments


**CareerZone - Myers Middle School**



During a press conference at the Stephen and Harriet Myers Middle School, this seventh grader...

All rights reserved  
Uploaded on Jun 28, 2010  
0 comments

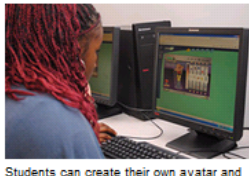
**CareerZone - Myers Middle School**



This young girl is ready to embark on a "Stemventure" to learn about possible careers in Science...

All rights reserved  
Uploaded on Jun 28, 2010

**CareerZone - Myers Middle School**



Students can create their own avatar and then send it through an exploration of STEM careers.

All rights reserved  
Uploaded on Jun 28, 2010

**CareerZone - Myers Middle...**

10 photos  
27 views

**Binghamton Mets and Department...**

8 photos  
50 views

**Worker Misclassification: It's...**

7 photos  
79 views

**The True Meaning of Memorial...**

18 photos  
213 views

**Labor Department Promotes...**

10 photos  
28 views

**Dr. King Career Fair**

14 photos  
16 views

http://worldcat.org/arcviewer/4/AO%23/2011/01/03/H1294085203555/viewer/extlink.html?link=http://www.flickr.c

Internet 75%

start New York (State... Z:\OCE\Archive... Microsoft Power... Flickr: NYSLabor'... 10:50 AM

# Twitter

twitter

Have an account? [Sign In](#)

## 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**.

[Sign Up](#)

Get updates via SMS by texting **follow NYSenate** to **40404** in the United States  
Codes for other countries



**NYSenate**

RT @NYSenateCIO: Announcing @NYSenate's OpenLegislation 1.6 "ember eltanin" <http://bit.ly/h0HGQz> #OpenNY #OpenGov #Gov20

11:04 AM Dec 22nd via TweetDeck

2nd meeting of Senate Leg Task Force on Demographic Research & Reapportionment under way. Live vid: <http://bit.ly/flWoLq> #redistricting

9:08 AM Dec 16th via TweetDeck

Uncut video of Tuesday's public hearing on #redistricting: <http://bit.ly/hnyoZg> #NY #NYS

8:41 AM Dec 16th via TweetDeck

Clear message of reform resonates from 1st scheduled mtg seeking public involvement in #redistricting process: <http://bit.ly/ifenoR>

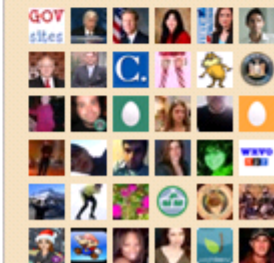
Name New York Senate  
Location Albany, NY  
Web <http://nysenate.gov>  
Bio News from the New York Senate.

235 following 4,979 followers 384 listed

Tweets 911

Favorites

Following



[View all...](#)

RSS feed of NYSenate's tweets

# Ideascale Blogs



The screenshot shows the header of the 'Straight Talk, Straight Answers' blog. The header features the 'taxpayer.ny.gov' logo on the left and the text 'Governor Paterson's Office of Taxpayer Accountability' and 'Straight Talk, Straight Answers' in the center. There are social media icons for Facebook and RSS on the right. Below the header is a navigation menu with links for HOME, STRAIGHT TALK FROM THE TAXPAYER, OFFICE OF TAXPAYER ACCOUNTABILITY, GOVERNOR PATERSON'S HOMEPAGE, and ABOUT. A search bar is located on the right side of the page. The main content area displays a blog post titled 'Straight Talk from the Taxpayer: Scheduling Online Payments' dated December 27, 2010. The post includes a quote from a taxpayer and a response from the office. A sidebar on the right contains a 'Thank You' message and information about the blog's purpose.

**taxpayer.ny.gov**

Governor Paterson's Office of Taxpayer Accountability

## Straight Talk, Straight Answers

HOME STRAIGHT TALK FROM THE TAXPAYER OFFICE OF TAXPAYER ACCOUNTABILITY GOVERNOR PATERSON'S HOMEPAGE ABOUT

"New post on Straight Talk, Straight Answers <http://bit.ly/9qpi8B> #nygov" – NYTaxpayer

Search SEARCH

### 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

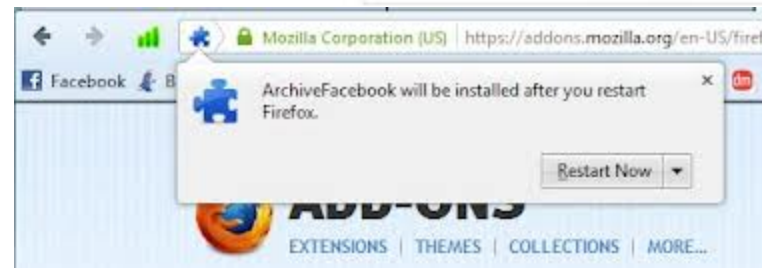
## ■ Facebook

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

## ■ ArchiveFacebook

- ❑ Firefox Add on

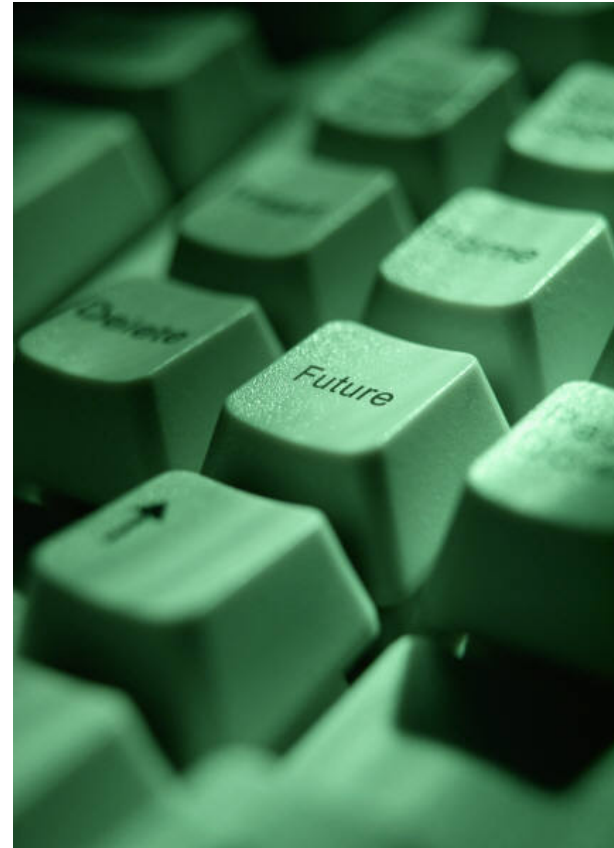
## ■ ArchiveSocial



---

# 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](mailto:mmartin@mail.nysed.gov)

---