



# A flickr of Hope: Harvesting Social Networking Sites with Archive-It

Best Practices Exchange 2009

September 2, 2009

# Web 2.0 and “open” government

- Politicians and government agencies are leveraging the outreach potential of web 2.0 applications
- Is content created using these tools a record?
- If so, how do we archive it?
  - Tweetake.com
  - Wordpress Lifestream plugin, [/](#)
  - Pasting comments in Outlook Archives account
  - ContextMiner
  - Archive-It

# Potential solutions



Desktop View  
download

- Downloads Content to the desktop
- Desktop view



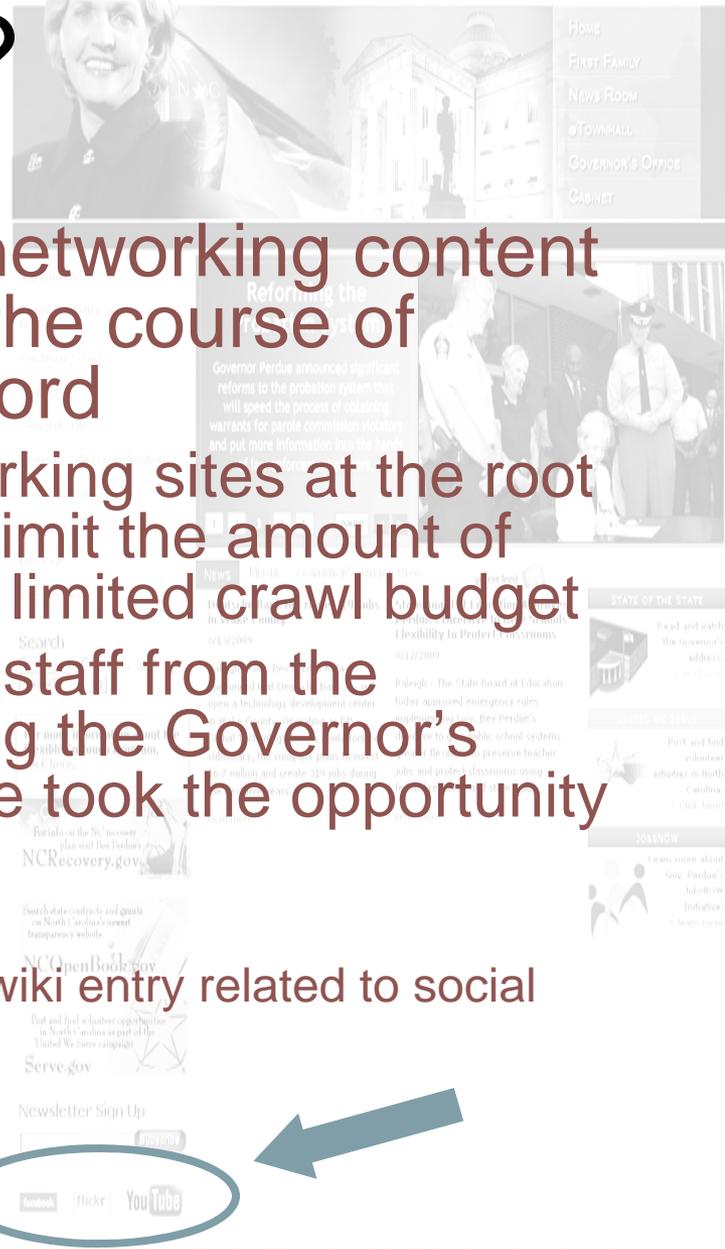
# What are we doing?

Partnering to pilot test the  
harvesting/preservation of  
state agency social networking  
content using Archive-It

# Why are we doing it?

We have always felt that social networking content created by state agency staff in the course of business is part of the public record

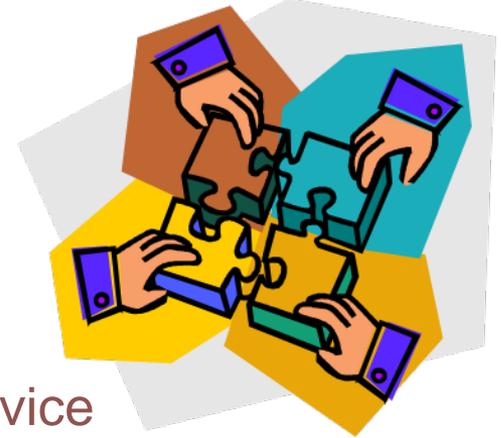
- But we constrained social networking sites at the root because we didn't know how to limit the amount of content captured and we have a limited crawl budget
- Recently, we were contacted by staff from the Governor's Office about capturing the Governor's social networking content and we took the opportunity to re-evaluate our approach
  - Reviewed options
    - Discovered Molly was working on wiki entry related to social networking site crawls



# What sites are we talking about?

- Currently working on flickr, Facebook, and twitter
- Next phase will be YouTube, MySpace, and LinkedIn (assuming no technological issues)
- Will continue to add new applications as we find out that NC agencies are using them

# Who is involved?



- Archive-It
  - Provides hosted subscription harvesting service
  - Identifies limitations of the crawler
  - Provides scoping help
- North Carolina Department of Cultural Resources
  - Manages the Archive-It subscription budget
  - Scopes crawls
  - Identifies agency seed sites
- North Carolina Governor's Office
  - Provides agency user perspective
  - Provides help in identifying agency social networking sites

# Lessons learned

- ☑ Check out Molly's wiki entry
- ☑ If you have multiple options, start with an easy one like flickr
- ☑ Always run a test crawl before you run a harvest
- ☑ Enlist help getting a listing of all "in scope" social networking sites

# The good news is...



- Archive-It is able to capture and render back flickr photostreams (including comments) very well
  - flickr how to's
    - Put the flickr photostream address in as your seed
      - <http://www.flickr.com/photos/bevperdue/>
    - Add a scoping rule if you want to capture the profile
      - [+http://\(com,flickr,www,\)/people/bevperdue](http://(com,flickr,www,)/people/bevperdue)
  - how big is it?
    - Our governor's flickr crawl returns about 12,500 documents (of course, this depends on the number of photos and comments in the photostream).

# flickr favorites on live web



flickr

You aren't signed in [Sign In](#) [Help](#)

[Home](#) [The Tour](#) [Sign Up](#) [Explore](#) ▾

[Search](#) ▾

**Governor Bev Perdue's photostream** [pro](#)

[Sets](#) [Tags](#) [Archives](#) [Favorites](#) [Profile](#)

Governor Bev Perdue doesn't have any favorites available to you.

Here's a link back to [Governor Bev Perdue's photostream](#).

You [Sign in](#) | [Create Your Free Account](#) [Bookmark on Delicious](#)

Explore [Places](#) | [Last 7 Days](#) | [This Month](#) | [Popular Tags](#) | [The Commons](#) | [Creative Commons](#) | [Search](#)

Help [Community Guidelines](#) | [The Help Forum](#) | [FAQ](#) | [Sitemap](#) | [Get Help](#)

[Flickr Blog](#) | [About Flickr](#) | [Terms of Use](#) | [Your Privacy](#) | [Copyright/MP Policy](#) | [Report Abuse](#)

[繁體中文](#) | [Deutsch](#) | [English](#) | [Español](#) | [Français](#) | [한국어](#) | [Italiano](#) | [Português](#)

Copyright © 2009 Yahoo! Inc. All rights reserved.

# flickr favorites harvested using Archive-It



You are viewing an archived Web site, archived on 0:26:09 Jul 22, 2009, that is part of a collection of archived websites created using [Archive-It](#). The information on this web page may be out of date. External links, forms, and search boxes may not function within this collection. [\[ hide \]](#)

flickr

You aren't signed in [Sign In](#) [Help](#)

[Home](#) [The Tour](#) [Sign Up](#) [Explore](#) ▾

[Search](#) ▾

**Governor Bev Perdue's photostream**

[Sets](#) [Tags](#) [Archives](#) [Favorites](#) [Profile](#)

Governor Bev Perdue doesn't have any favorites available to you.

Here's a link back to [Governor Bev Perdue's photostream](#).

You [Sign in](#) | [Create Your Free Account](#) [Bookmark on Delicious](#)

Explore [Places](#) | [Last 7 Days](#) | [This Month](#) | [Popular Tags](#) | [The Commons](#) | [Creative Commons](#) | [Search](#)

Help [Community Guidelines](#) | [The Help Forum](#) | [FAQ](#) | [Sitemap](#) | [Help by Email](#)

[Flickr Blog](#) | [About Flickr](#) | [Terms of Use](#) | [Your Privacy](#) | [Copyright/MP Policy](#) | [Report Abuse](#)

[繁體中文](#) | [Deutsch](#) | [English](#) | [Español](#) | [Français](#) | [한국어](#) | [Italiano](#) | [Português](#)

Copyright © 2009 Yahoo! Inc. All rights reserved.

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.

# is a bit trickier

- Archive-It can capture and render back Facebook content, but
  - the content is not on a single URL so you will need to add scoping rules to capture content on other URLs (notes and photo albums fall into this category),
  - some portions are blocked by a robots.txt script so we have requested that Facebook work with us on this issue (photo albums fall into this category), and
  - there is a lot of Javascript and to get that to render best requires the users to turn off Javascript in their browser and to be in “proxy-mode” which is something our networks in DCR don’t allow so Archive-It is working on another viewing option which is still in beta

# Facebook album on live web



# Facebook album harvested using Archive-It



facebook

Remember Me  Forgot your password?

Facebook helps you connect and share with the people in your life.



North Carolina State Government Web Site Archive Web Archive (North Carolina State Archives and State Library of North Carolina)



Enter Web Address:

## Bev Perdue's Photos - First Day in Office

Photo 6 of 6 | [Back to Album](#) | [Bev Perdue's Photos](#) | [Bev Perdue's Profile](#)

[Previous](#) [Next](#)



Cabinet Swearing-In Ceremony

From the album:  
"First Day in Office" by Bev Perdue

Added February 2

## Not in Archive

The page you requested has not been archived.

Most likely the page you are requesting was outside of the crawlers scope. Try another request or click [here](#) to search for all pages on the same host as [www.facebook.com](http://www.facebook.com).

Or this error message could have appeared because the site is currently being crawled and the archived pages are not available in the Wayback Machine yet. It usually takes about an hour after your crawl has finished for the site to appear in the Wayback Machine. Please try again after the allotted time, and if you continue to see this error for a page you believe to have been crawled, [contact us](#).

You can also try searching for [www.facebook.com/album.php?aid=95395&id=11552180685](http://www.facebook.com/album.php?aid=95395&id=11552180685) on the [live web](#) or in the [global Wayback Service](#).

[Home](#) | [Copyright © 2005, Internet Archive](#) | [Terms of Use](#) | [Privacy Policy](#)

# Facebook made easy



- Facebook how to's
  - Put the Facebook account address in as your seed
    - <http://www.facebook.com/pages/Bev-Perdue/11552180685/>
  - Add scoping rules for any content on an address other than the account address (notes, photos, videos, etc.)
    - `+http://(com,facebook,www,)/note`
    - `+http://(com,facebook,www,)/album`
- how big is it?
  - Our governor's Facebook crawl returns about 1,800 documents (of course, this depends on the number of notes, photos, and comments).

# So what about twitter?



From <http://www.acriddle.com/>

- Capturing just the tweets of the agency is a breeze, but trying to capture the links in the tweets or the entire conversation is a bit harder
  - twitter how to's
    - Put the twitter address in as your seed (make sure you do not include www)
      - <http://twitter.com/VisitNCWine/>
  - twitter tips
    - Be aware that many tweets include links and determine how you want to handle them (in or out of scope).
    - If you are missing content like backgrounds or images, make sure you are not constraining a host that is serving this content, like [3s.amazonaws.com](http://3s.amazonaws.com)
    - Our VisitNCWine twitter crawl returns about 120 documents without one hop off and 2000 with one hop off turned on.

twitter on live web



twitter harvested using Archive-It  
(One hop off feature turned off)



twitter Login Join Twitter

Hey there! VisitNCWine is using Twitter.

Twitter is a free service that lets you keep in touch with people through the exchange of quick, frequent answers to one simple question: What are you doing? **Join today** to start receiving VisitNCWine's tweets.

**Join today!**

Already using Twitter from your phone? [Click here.](#)

---

**VisitNCWine**

StarNews featured muscadine wine, along with @DuplinWinery and Bannerman Vineyard. <http://bit.ly/17YRVL>

1:32 PM Aug 13th from TweetDeck

---

Iron Gate Vineyards Dixie Dawn included in the winning menu for Best Dish in NC Casual Dining! <http://bit.ly/1Johz2N>

7:17 AM Aug 13th from TweetDeck

---

@ncwinetv reviews McRitchie Winery and Ciderworks Hard Cider. <http://bit.ly/1lmzQ>

2:01 PM Aug 12th from TweetDeck

---

The Atlanta Journal-Constitution recommends touring Raffaldini Vineyards. <http://bit.ly/v2cvK>

7:18 AM Aug 12th from TweetDeck

---

Wondering what to make for dinner tonight? Try this trout recipe paired with @ChilodressWines Pinot Gris. <http://bit.ly/tc4QA>

1:09 PM Aug 11th from TweetDeck

---

@DuplinWinery swallowing higher tax. <http://bit.ly/1QTN5b>

8:54 AM Aug 11th from TweetDeck

---

Winston-Salem Journal enjoyed a wine day trip to Shelton Vineyards, McRitchie Winery and Raffaldini Vineyards. <http://bit.ly/WixP6>

12:37 PM Aug 10th from TweetDeck

---

Asheville Citizen-Times profiles the thriving North Carolina wine

**Name** NC Wine  
**Location** North Carolina  
**Web** <http://www.visitncwine.com>  
**Bio** NC Wine & Grape Council enhances product quality for consumers, economic viability & opportunity for growers & processors thru education, marketing & research

60 128  
following followers

**Tweets** 29

**Favorites**

**Following**

[View All...](#)

**RSS feed of VisitNCWine's tweets**

You are viewing an archived Web site, archived on 11:37:01 Aug 26, 2009, that is part of a collection of archived websites created using Archive-It. The information on this web page may be out of date. External links, forms, and search boxes may not function within this collection. [\[hide\]](#)

twitter Login Join Twitter

Hey there! VisitNCWine is using Twitter.

Twitter is a free service that lets you keep in touch with people through the exchange of quick, frequent answers to one simple question: What are you doing? **Join today** to start receiving VisitNCWine's tweets.

**Join today!**

Already using Twitter from your phone? [Click here.](#)

---

**VisitNCWine**

Our State magazine features Yadkin Valley, Hanover Park, RagApple Lassie, RayLen, Round Peak, Shelton, Stony Knoll, Surry CC & Westbend!

about 22 hours ago from TweetDeck

---

The Times-News shares the award-winning recipe for the NC Best Dish featuring Iron Gate Dixie Dawn. <http://bit.ly/Gh1pu>

7:20 AM Aug 24th from TweetDeck

---

Wine down and see what wine events are happening this weekend! [http://www.visitncwine.com/...](http://www.visitncwine.com/)

11:19 AM Aug 21st from TweetDeck

---

The Examiner previews the Pinehurst Food and Wine Festival on Labor Day weekend. <http://bit.ly/18CYAX>

6:35 AM Aug 21st from TweetDeck

---

Raffaldini wines will be featured in a five-course wine dinner at Durham's Four Square Restaurant. <http://bit.ly/3KirQG>

1:32 PM Aug 19th from TweetDeck

---

Metro Magazine highlights Raffaldini, @DuplinWinery, and Haw River Valley AVA and its wineries. <http://bit.ly/h56en>

7:53 AM Aug 19th from TweetDeck

**Name** NC Wine  
**Location** North Carolina  
**Web** <http://www.visitncwine.com>  
**Bio** NC Wine & Grape Council enhances product quality for consumers, economic viability & opportunity for growers & processors thru education, marketing & research

60 159  
following followers

**Tweets** 38

**Favorites**

**Following**

[View All...](#)

**RSS feed of VisitNCWine's tweets**

## twitter following/followers/favorites on live web



The screenshot shows the Twitter sign-in page on a live web browser. The Twitter logo is in the top left. In the top right, there are links for "Login" and "Join Twitter!". The main content area is split into two columns. The left column is titled "Sign in to Twitter" and contains a sub-header: "If you've been using Twitter from your phone, [click here](#) and we'll get you signed up on the web." Below this are two input fields: "Username or email" and "Password". To the right of the password field is a link "Forgot?". Below the password field is a checkbox labeled "Remember me" and a "Sign In" button. The right column is titled "Create Your Account" and features a blue "Join!" button. Below the button is the text "Already using Twitter from your phone? [Click here.](#)" and a "Select Language ..." dropdown menu. At the bottom of the page is a footer with copyright information and various links: "© 2009 Twitter About Us Contact Blog Status Goodies API Business Help Jobs Terms Privacy".

## twitter following/followers/favorites harvested using Archive-It (One hop off feature turned on)



You are viewing an archived Web site, archived on 17:59:45 Aug 14, 2009, that is part of a collection of archived websites created using [Archive-It](#). The information on this web page may be out of date. External links, forms, and search boxes may not function within this collection. [[hide](#)]

The screenshot shows the Twitter sign-in page as it appeared in an Archive-It archive. The layout is identical to the live web version, but with a yellow banner at the top containing the Archive-It notice. The Twitter logo and navigation links are present. The sign-in form on the left and the "Create Your Account" section on the right are visible. The footer at the bottom contains the same copyright and link information as the live version.

# Sample captures

- <http://www.archive-it.org/collections/194>
  - <http://twitter.com/>
  - <http://www.facebook.com/>
  - <http://www.flickr.com/>



# Next steps



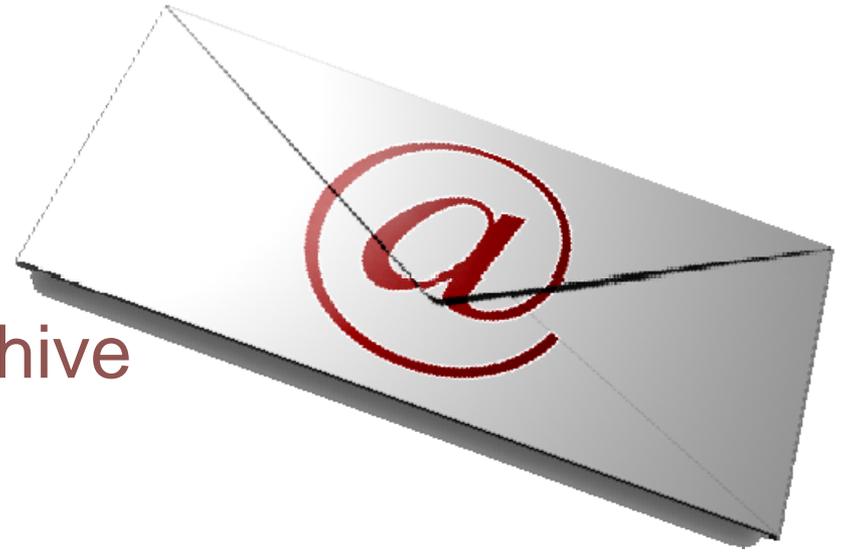
- Deal with need for “proxy-mode” viewing (Archive-It & DCR)
- Deal with Facebook robots.txt script (ALL)
- Gather list of existing social networking sites that are part of the public record (DCR & Gov’s Office)
- Work on capturing content on other sites being used by agencies like MySpace, LinkedIn, and YouTube (Archive-It & DCR)
- Determine how often to capture based on our funding and subscription budget constraints (DCR)

# Questions for you...

- Are others capturing this type of content?
- Have you been successful?
- What tools are you using?
- What sites have you been capturing?
- What issues are you running into?
- What agencies are you partnering with?
- Has anyone had any success partnering with social networking sites to facilitate harvests?

**Do you have questions for us?**

# Contact Info.



- Molly Bragg, Internet Archive  
[mbragg@archive.org](mailto:mbragg@archive.org)
- Kelly Eubank, North Carolina State Archives  
[kelly.eubank@ncdcr.gov](mailto:kelly.eubank@ncdcr.gov)
- Jennifer Ricker, State Library of North Carolina  
[jennifer.ricker@ncdcr.gov](mailto:jennifer.ricker@ncdcr.gov)