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

- 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
  - 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).
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 notes on live web
Facebook notes harvested using Archive-It

Facebook photos on live web
Facebook photos harvested using Archive-It

Facebook info on live web
Facebook info harvested using Archive-It

Facebook album - Bev Perdue’s Photos - First Day in Office

Previous: Next

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.

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.

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?

- 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
    - 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)
You are viewing an archived website, archived on 17 Feb 15-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]
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

- Kelly Eubank, North Carolina State Archives
  kelly.eubank@ncdcr.gov

- Jennifer Ricker, State Library of North Carolina
  jennifer.ricker@ncdcr.gov