

Assessing and Analyzing Data: Establishing Policy and Programs

Kelly Eubank, North Carolina State Archives
Jennifer Ricker, North Carolina State Library

Who, What, Why & How Long

- Jennifer Ricker, Digital Collections Manager, State Library of NC
- Kelly Eubank, Electronic Records Archivist, North Carolina State Archives
- Publications and Records—all media, governed by statute
- Pilot Partner-Sept. 2005

Archive-It Pilot Scope

- 3 collections of up to 100 websites each
- 10 million documents total, up to .5 terabytes of data

Scope Continued

- Crawled 3 collections—Cabinet, Council of State, Boards and Commissions
- 274 web addresses
- Initial search was daily, scaled to weekly after 3 weeks.
- End of Project— 9.5 million objects
- Archive-it captured 54 different formats

Lessons Learned For IA

- One hop off
 - “out of scope materials”
 - “inappropriate materials”
 - Search for “lottery”
 - Tremendous amount of material would have to be masked.
- Re-do analysis to go into production
 - Turn off of one hop off feature.
- Cannot search across collections
 - Clumsy search
 - Broken links

North Carolina State Government Web Site Archives



North Carolina State Government Web Site Archives

North Carolina State Archives State Library of North Carolina

Search the Web Archives

Search

[Search Help](#)

Welcome!

The *North Carolina State Government Web Site Archives* allows you to view North Carolina state agency web sites from past dates. The *Web Site Archives* contains web sites from the Fall of 2005, and from April 2006 forward allowing free and open access to this information long after the sites have changed on the live web.

The *Web Site Archives* can be searched via the search box on the top of every page on this site. For tips and helpful hints on searching, read our [Search Help](#) document. Users may also browse for web sites by [State Agency](#) or by [Collection](#).

The *Web Site Archives* began as a pilot project with the [Internet Archive](#) during the Fall of 2005. The purpose of this project was to refine a tool called [Archive-It](#) which collects, preserves, and provides access to web sites of enduring value.

The success of the Pilot Project led to the creation of the current version of *The North Carolina State Government Web Site Archives* which began archiving web sites in late April of 2006. The *Web Site Archives* contains copies of state agency web sites captured during the pilot project in the Fall of 2005 and after the official launch in April 2006 forward.

The *North Carolina State Government Web Site Archives* is proof of the ongoing commitment by the North Carolina State Archives and the State Library of North Carolina to provide free and open access to state government records and publications.

This web page has been visited:
000017 times since October 19, 2006.



Above: Screenshot of the Governor's web site from Sept. 20th, 2005. Click image to view site

Program Document

North Carolina Department of Cultural Resources

Program for Maintaining and Preserving Records of Web-Based Activities

Introduction

In North Carolina, as in other governments around the country, the World Wide Web has become the preferred method for state agencies, county governments, municipalities, and other governmental agencies to disseminate information, provide services, and transact business with its citizenry. Much of the information posted on Web sites by government agencies exists only in electronic format and is not available through other means. The ever increasing use of Web sites by North Carolina's government agencies complicates the wide spectrum of electronic records management issues facing government agencies including: storage, preservation, access, and authenticity. The identification, selection, capture, and preservation of government Web sites is sanctioned under the Digital Preservation Policy Framework and approved by the Department of Cultural Resources (DCR).

Legal Background

State government Web sites have unique identities; they publish information that

Standards Document

Determination of Which Web sites Should be Captured

STEP I: In Scope or Out of Scope

1. All Web sites selected for inclusion are sites that contain official state government information, either hosted on government Web servers or hosted by private companies working for the agency.

Example:

http://www.ncbar.com	North Carolina State Bar	State Government Web site, hosted on a state server
http://www.ncartmuseum.org/	North Carolina Museum of Art	State Government Web site, hosted on a private server

2. Private non-profit organizations and for-profit organizations are NOT collected, even if they assist some government agency.

Example:

http://www.nczoo.com/	North Carolina Zoological Society	Nonprofit organization supports zoo activities, but is separate from the NC Zoo
---	-----------------------------------	---

Macro-Appraisal Table

	1 point – Low value	2 points – Medium value	3 points – High value
1: Size	Small Web site with few directories	Large Web site with multiple directories	Large Web site with multiple directories from multiple offices
2: Originality	Mainly copies of paper publications	Combination of original material and copies of paper material	Substantially original material not available in paper form
3: Frequency of update	Static Web site with time-insensitive material; information rarely changes	Web site updated on a regular basis; information current and fresh	Web site information changes frequently; timeliness and currency of information extremely important
4: Historical value	Information is ephemeral; little value beyond the present day	Some original information with historical value; partially documents decisions/trends of government and programs	Rich original information; statistics/facts document government decisions/trends and programs
5: Evidential value	Information never litigated and little or no legal risk	Potential for litigation	Substantial liability
6: Public interest	Low public interest; little press coverage	Medium public interest; some press coverage; potentially controversial information	High public interest; frequent press coverage; public interest groups monitor sites
7: Governmental interest	Little legislative, executive, or judicial interest	Some legislative, executive or judicial interest	Active and vigilant legislative, executive, or judicial interest

Procedures Document

Selection Criteria

Limitations of Archive-It

Even if a Web site falls “In Scope” according to the DCR’s Web site Capture Standards, it may fall “Out of Scope” in Archive-It. The following are examples where technical limitations, space considerations, and funding limitations preclude capture in Archive-It.

- Due to technological, procedural, and funding limitations the Web sites of local governments and public colleges and universities are NOT collected at this time, although the North Carolina Community College System Office is collected as a state agency.

Example:

http://www.waketech.edu/	Wake Technical Community College	Education Web site
http://www.wakegov.com/	Wake County Government	Local North Carolina government site

- Web sites which consist entirely of Web-enabled databases or which require user input to access information are excluded, even if they meet other criteria, because the capture software currently in use is not capable of accessing the information. Whenever such technical limitations arise, if the Web site or Web sites are deemed to be appropriate for

Master Seed List with Ranking

	A	K	L	Q	R
1	Domain	Total	Agency	Lowest Level	Notes
98	www.investnc.com	12	Commerce		
99	www.iso.scio.nc.gov	13.6666667	Governor	ITS - Enterprise Security and Risk Mgmt.	sidebar: site indicates that they are responsible for classifying information according to state law.
100	www.itpi.dpi.state.nc.us	8.5	DPI	Technology Planning and Support	Web Resources for North Carolina Educators
101	www.its.state.nc.us	14	Governor		
102	www.its.state.nc.us/ITProcurement/	12	Governor	ITS - State IT Procurement	
103	www.jfk.adaco.net	9	DHHS	Julian F. Keith Alcohol and Drug Abuse Treatment Center,	
104	www.ltgov.state.nc.us	12.25	Lt. Governor		
105	www.mro.enr.state.nc.us	12	DENR	Mooresville Regional Office	
106	www.murdochcenter.org	9.5	DHHS		
107	www.myeatSMARTmove.com	8.25	DHHS		KM - New domain?
108	www.naturalsciences.org	11.5	DENR		
109	www.nc.ngb.army.mil	13.5	CCPS		NC National Guard
110	www.nc5aday.com	8.5	DHHS	Public Health Division	

Verification Steps

c) Decisions regarding material previously blocked by robots should be noted in the note section of the **Distribution Robots Report**.

In addition, for any material that a team member decides to not pursue, the team member should note their reasoning and send email to the other members for review.

2) Redirects

a) Use the **Seed Status Report** to see what was redirected, timed out, etc. If a site is unreachable, the team should be notified so that it can make a decision whether or not to disable the seed.

b) This is a clean up exercise to ensure that we have to correct seeds. Redirects should be left alone unless the previous URL no longer functions for that site. If we disable the seed that redirects, we will be unable to hook the new seed to previously captured instances of the page.

3) Identifying out of scope URLs to block access.

a) Review the **Seed Source Report** for out of scope material. **The Seed Source Report** contains a Host column. Look at the Host column and check the URL on the live Web to verify that it is out of scope.

b) Alert the team regarding proposed “out of scope” hosts and get their feedback.

c) Once the team agrees that the host is out of scope and agrees to disable it, click into Archive-It, choose the correct collection, and click the “Crawl Management” tab. Then click on “Host Constraints.” Enter the host to be blocked click add, then check the “Block Completely” box and click “Update constraints.” These steps should block all out of scope hosts.

Content Not Being Captured

The screenshot shows a Mozilla Firefox browser window with the title "Archive-It/Internet Archive Wayback Machine - Mozilla Firefox". The address bar contains the URL "http://wayback.archive-it.org/195/20050922190821rn_1/www.ncstatefair.org/2005/". The page content includes the Internet Archive logo, the text "North Carolina Council of State Web Web Archive (North Carolina State Archives)", and the Wayback Machine logo. A search bar is present with the text "Enter Web Address: http://". Below the search bar, the text "Robots.txt Retrieval Exclusion." is displayed. The main message reads: "We're sorry, access to <http://www.ncstatefair.org/2005/> has been blocked by the site owner via robots.txt. [Read more about robots.txt](#). [See the site's robots.txt file.](#) Try another request or click here to search for all pages on [ncstatefair.org/2005/](http://www.ncstatefair.org/2005/). See the [FAQs](#) for more info and help, or [contact](#) us." At the bottom of the page, there are links for "Home", "Copyright © 2005, Internet Archive", "Terms of Use", and "Privacy Policy". The Windows taskbar at the bottom shows the Start button, several open applications (Calculator, Microsoft Office Word, Firefox, Microsoft Office PowerPoint, Windows Explorer), and the system tray with the time 9:36 AM.

Archive-It/Internet Archive Wayback Machine - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://wayback.archive-it.org/195/20050922190821rn_1/www.ncstatefair.org/2005/

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

ARCHIVE-IT

North Carolina Council of State Web Web Archive (North Carolina State Archives)

INTERNET ARCHIVE
WaybackMachine

Enter Web Address: http:// All Take Me Back Adv. Search

Robots.txt Retrieval Exclusion.

We're sorry, access to <http://www.ncstatefair.org/2005/> has been blocked by the site owner via robots.txt.
[Read more about robots.txt](#)
[See the site's robots.txt file.](#)
Try another request or click here to search for all pages on [ncstatefair.org/2005/](http://www.ncstatefair.org/2005/).
See the [FAQs](#) for more info and help, or [contact](#) us.

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

Done

start Calculator 5 Microsoft ... 8 Firefox 4 Microsoft ... 2 Windows E... Microsoft Pow... 9:36 AM

You are viewing an archived Web site, archived on 18:41:04 Nov 02, 2006, that is part of a collection of archived websites created using [Archive-It](#). The content may not be up to date. External links, forms, and search boxes may not function within this collection. [[hide](#)]

2006 State Fair
FEED YOUR SENSES

[Home](#) |
 [General Info](#) |
 [Tickets](#) |
 [Entertainment](#) |
 [Exhibits](#) |
 [Competitions](#) |
 [Newsroom](#) |
 [Contact Us](#)

[Fairgrounds Events Calendar](#)
[Fairgrounds Facilities & Rental Rates \(non-fair\)](#)
[Buy the book - The N.C. State Fair: The First 150 Years](#)
[Fairgrounds Map: Non-Fair pdf | Fair-time pdf](#)
[Driving Directions](#)
[Member of the International Assn. of Fairs & Expositions](#)
[N. C. County Fairs](#)
[Frequently Asked Questions](#)
[News Releases](#)

The sun has set on the 2006 State Fair

State Fair brings in 785,956 fairgoers for 5th highest attendance ever
Come back next year, Oct. 12-21, 2007!

[Check out this aerial photo of the Fair!](#)



Daily Schedules

[Friday, Oct. 13](#)
[Saturday, Oct. 14](#)
[Sunday, Oct. 15](#)
[Monday, Oct. 16](#)
[Tuesday, Oct. 17](#)
[Wednesday, Oct. 18](#)
[Thursday, Oct. 19](#)
[Friday, Oct. 20](#)
[Saturday, Oct. 21](#)
[Sunday, Oct. 22](#)

Need New Recipes?
[Check out the winners from our special cooking contests](#)

Special th
spo





october 12-21

Home | General Info | Tickets | Entertainment | Exhibits | Competitions | Sponsors | Newsroom



- Grounds Facilities & Rental Info
- IAFE
- NC County Fairs
- '07 Fairgrounds Events Calendar
- Vendors

Seriously Twisted Fun Coming Oct. 12-21, 2007

advance tickets on sale

2007 State Fair RALEIGH, NC

sept. 27 - oct. 11

Buy in advance and **save!**

BUY TICKETS

Concert tickets are on sale now!

2007 State Fair RALEIGH, NC

SUBWAY

CAN YOU BUILD A SERIOUSLY TWISTED SUB?

MEET! MAKE AS BE HASTS TWISTED TUESDAY

A YEAR OF SUBS | SUBWAY SOUP | TICKETS

ENTER NOW

Win a year's worth of free Subway

THE ORIGINAL FARM ANIMAL FRENZY

The farm animals are taking over

using Archive-It. T