A Web Crawl Is Not a Web Crawl, Is Not a Web Crawl

Tools for the Capture of Digital Assets on Websites

October 27, 2006

Kelly Eubank and Jennifer Ricker

Why Capture Websites?

- Websites now the primary way that North Carolina state agencies communicate with the public
- Over 80% of publications disseminated through the Web
- Important records on the Web:
  - Minutes, speeches (video), policies, images
- Websites have become an important part of agency history

What is Web Capture?

- Web crawler or "spider" collects web content
- Starts at predetermined list of URLs
- Makes a copy of web page, including all objects that are part of the web page
- Follows hyperlinks and captures additional web pages, as long as part of acceptable domain list
- Content captured is "clickable content"
  - must be a link for spider to find
  - must not require input from user

Website Philosophy

- Arizona Model vs. “Newspaper” theory
  - Arizona
    - Capture identified “series” hosted on website
    - Crawler only goes to the specified site
  - Newspaper
    - Website contains information that might be valuable or interesting to future research

Website Archiving Activities to Date

- Website guidelines and capture from 2001
- http://www.ah.dcr.state.nc.us/records/e_records/default.htm#web
- CEP from IMLS grant
- WAW from OCLC, NDIIPP partnership
- Archive-It from Internet Archive

Capturing Electronic Publications

- Developed by the University of Illinois through an IMLS grant
- From September 2004-2005, NC State Library participated in a pilot program to capture websites.
- Open-source software
- 200 state government sites, monthly
- Currently, 273 gigabytes
- No search mechanism

Web Archives Workbench

- Developed by OCLC through National Digital Information Infrastructure Preservation program
- Based on the Arizona Model for Web site Harvesting
- Consists of 4 tools: Discovery tool, Properties tool, Analysis tool, Harvest and Packager tools
- Beta Testing phase—to be completed in 2008

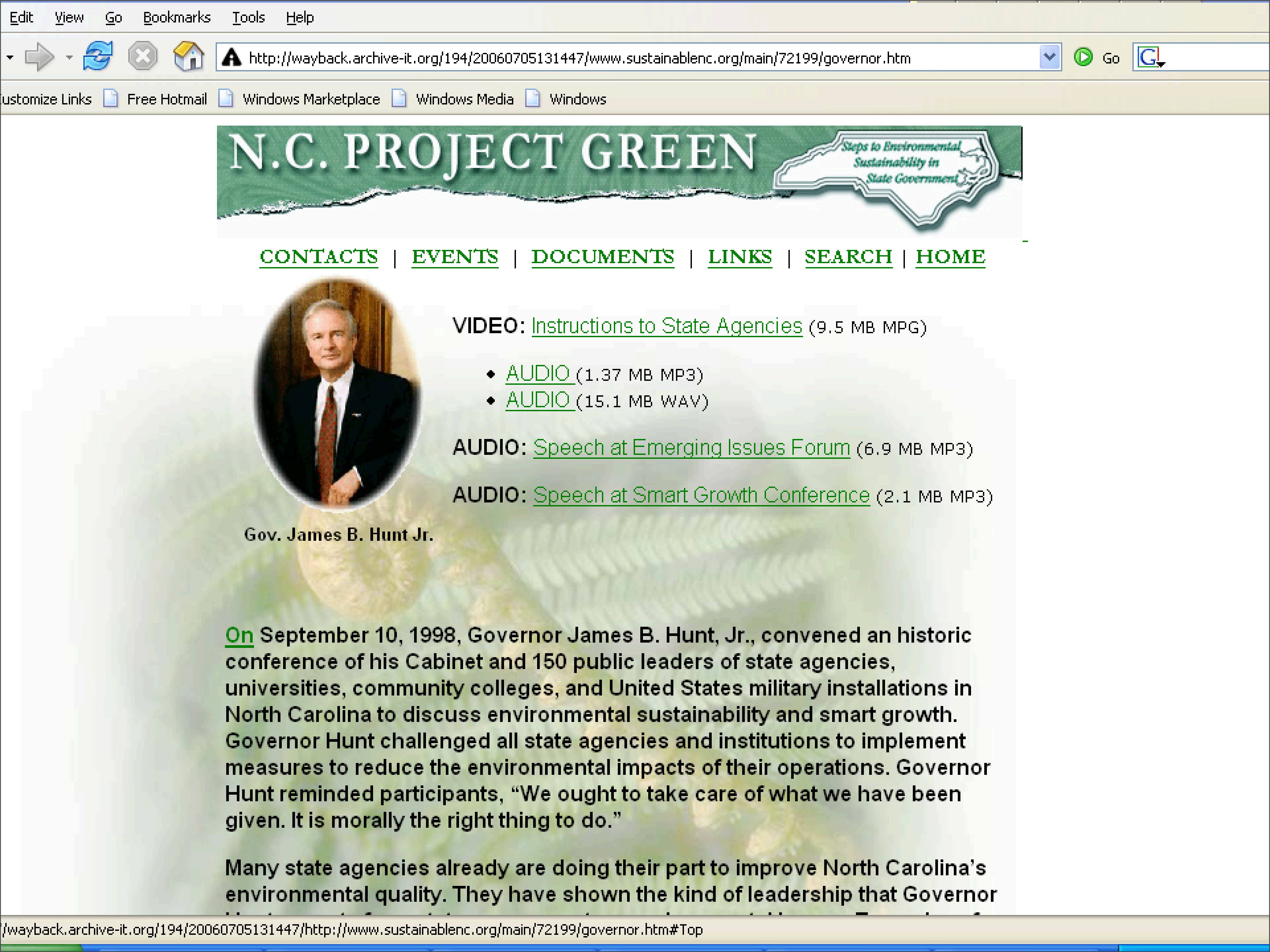
Archive-It

- Pilot: 3 collections of 100 websites each
- 10 million documents total, up to .5 terabytes of data
- Crawled 3 collections—Cabinet, Council of State, Boards and Commissions
- End of Project– 9.5 million objects
- Archive-it captured 54 different formats

ARCHIVING THE INTERNET FOR FUTURE GENERATIONS
COLLECT IT, MANAGE IT, SEARCH IT....ARCHIVE-IT

Discussion of Capture: What they can capture

- The Web contains valuable archival records in varying file formats.
  - Minutes—Microsoft Word, PDF
  - Speeches--Streaming Video, MP3, .WAV
  - Images--.JPEG, .GIF, .PNG
  - Policies--.HTML, .PDF
- CEP-Ability to capture but limited access
- WAW designed to capture documents
- Archive-It designed to capture the entire web site and render back


Edit
View
Go
Bookmarks
Tools
Help
http://wayback.archive-it.org/194/20060705131447/www.sustainablenc.org/main/72199/governor.htm
Go
ustomize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
Steps to Environmental Sustainability in State Government
/wayback.archive-it.org/194/20060705131447/http://www.sustainablenc.org/main/72199/governor.htm#Top

N.C. PROJECT GREEN

CONTACTS | EVENTS | DOCUMENTS | LINKS | SEARCH | HOME

Gov. James B. Hunt Jr.

VIDEO: Instructions to State Agencies (9.5 MB MPG)

- AUDIO (1.37 MB MP3)
- AUDIO (15.1 MB WAV)

AUDIO: Speech at Emerging Issues Forum (6.9 MB MP3)

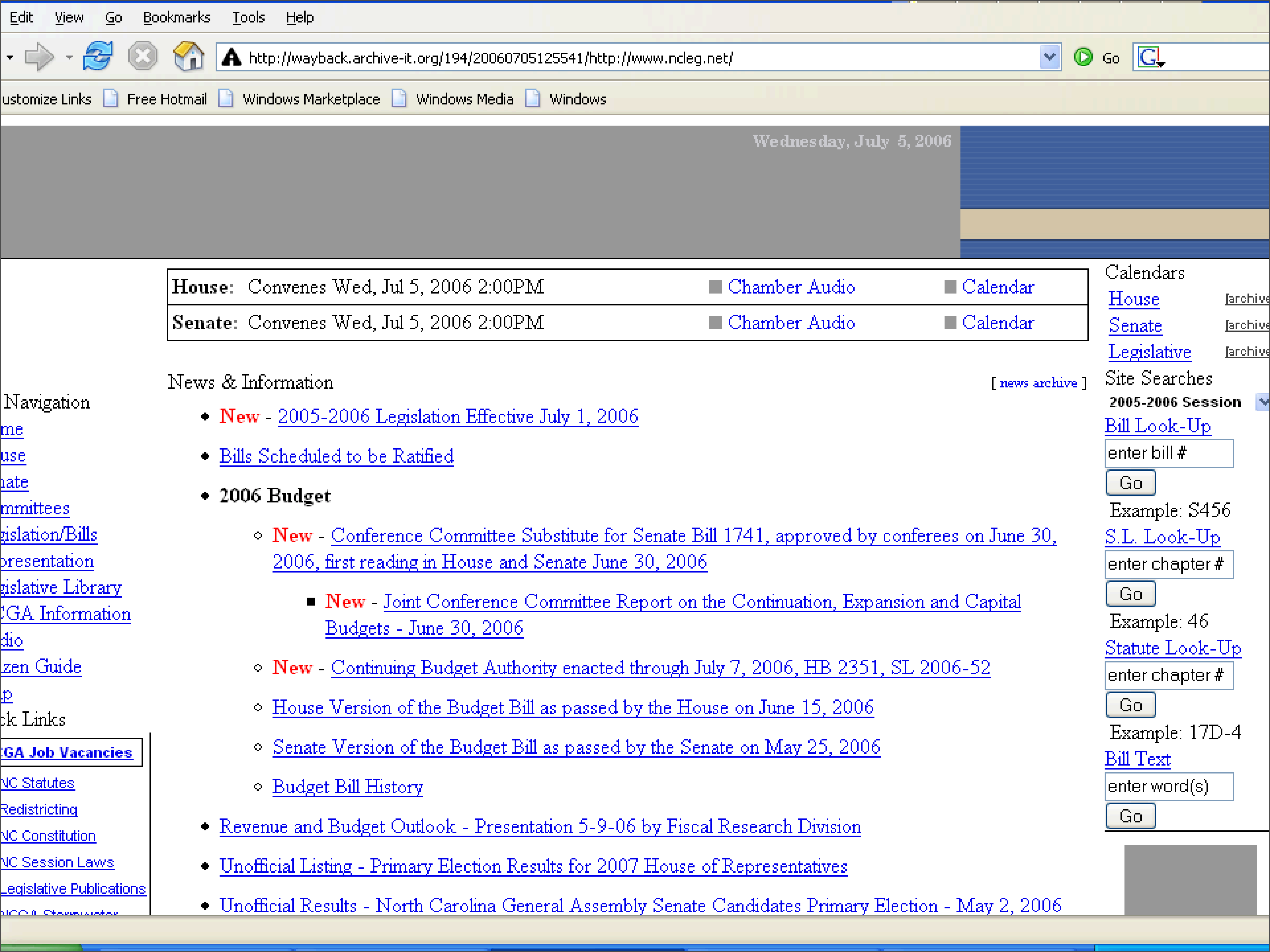
AUDIO: Speech at Smart Growth Conference (2.1 MB MP3)

On September 10, 1998, Governor James B. Hunt, Jr., convened an historic conference of his Cabinet and 150 public leaders of state agencies, universities, community colleges, and United States military installations in North Carolina to discuss environmental sustainability and smart growth. Governor Hunt challenged all state agencies and institutions to implement measures to reduce the environmental impacts of their operations. Governor Hunt reminded participants, “We ought to take care of what we have been given. It is morally the right thing to do.”

Many state agencies already are doing their part to improve North Carolina’s environmental quality. They have shown the kind of leadership that Governor

Discussion of Capture

- Inability to capture and display information
- Redirect to Live Web
  - Relative versus absolute links: (CEP)
  - **http://www.ncleg.net**
  - Java Script: (AI)
  - http://www.ncstatefair.org/
  - Community Colleges
  - http://www.ncccs.cc.nc.us/



http://wayback.archive-it.org/194/20060705125541/http://www.ncleg.net/

Wednesday, July 5, 2006

House:	Convenes Wed, Jul 5, 2006 2:00PM	■ Chamber Audio	■ Calendar
Senate:	Convenes Wed, Jul 5, 2006 2:00PM	■ Chamber Audio	■ Calendar

Navigation

me
use
nate
mmittees
gislation/Bills
presentation
gislative Library
GA Information
dio
zen Guide
lp

ck Links

GA Job Vacancies

NC Statutes
Redistricting
NC Constitution
NC Session Laws
Legislative Publications
NCGA Stormwater

News & Information

[news archive]

- New - 2005-2006 Legislation Effective July 1, 2006
- Bills Scheduled to be Ratified
- **2006 Budget**
  - New - Conference Committee Substitute for Senate Bill 1741, approved by conferees on June 30, 2006, first reading in House and Senate June 30, 2006
    - New - Joint Conference Committee Report on the Continuation, Expansion and Capital Budgets - June 30, 2006
  - New - Continuing Budget Authority enacted through July 7, 2006, HB 2351, SL 2006-52
  - House Version of the Budget Bill as passed by the House on June 15, 2006
  - Senate Version of the Budget Bill as passed by the Senate on May 25, 2006
  - Budget Bill History
- Revenue and Budget Outlook - Presentation 5-9-06 by Fiscal Research Division
- Unofficial Listing - Primary Election Results for 2007 House of Representatives
- Unofficial Results - North Carolina General Assembly Senate Candidates Primary Election - May 2, 2006

Calendars

House [archive
Senate [archive
Legislative [archive

Site Searches

2005-2006 Session

Bill Look-Up

enter bill #

Go

Example: S456

S.L. Look-Up

enter chapter #

Go

Example: 46

Statute Look-Up

enter chapter #

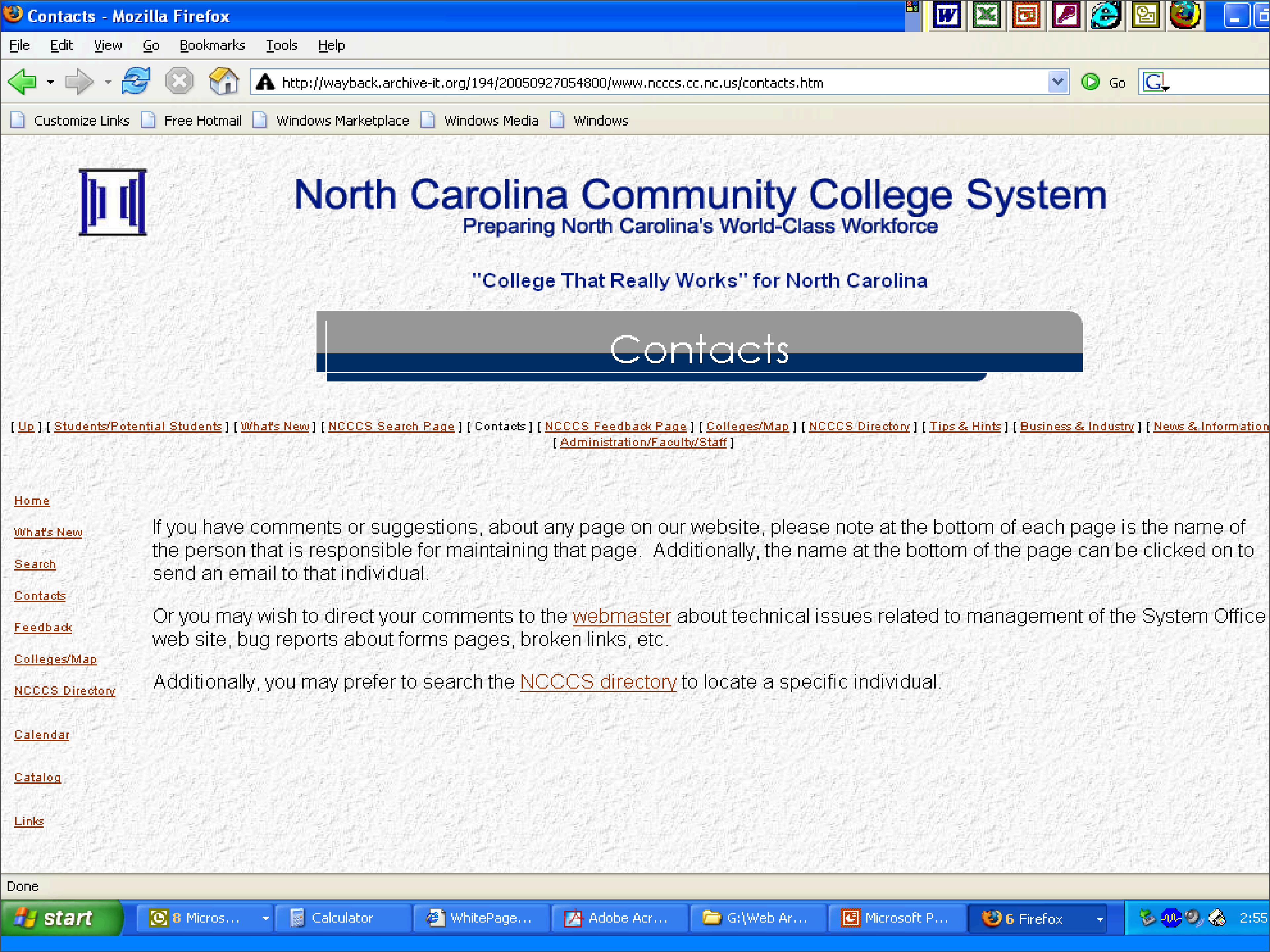
Go

Example: 17D-4

Bill Text

enter word(s)

Go



http://wayback.archive-it.org/194/20050927054800/www.ncccs.cc.nc.us/contacts.htm

North Carolina Community College System

Preparing North Carolina's World-Class Workforce

"College That Really Works" for North Carolina

Contacts

[Up] [Students/Potential Students] [What's New] [NCCCS Search Page] [Contacts] [NCCCS Feedback Page] [Colleges/Map] [NCCCS Directory] [Tips & Hints] [Business & Industry] [News & Information
[Administration/Faculty/Staff]

Home

What's New

Search

Contacts

Feedback

Colleges/Map

NCCCS Directory

Calendar

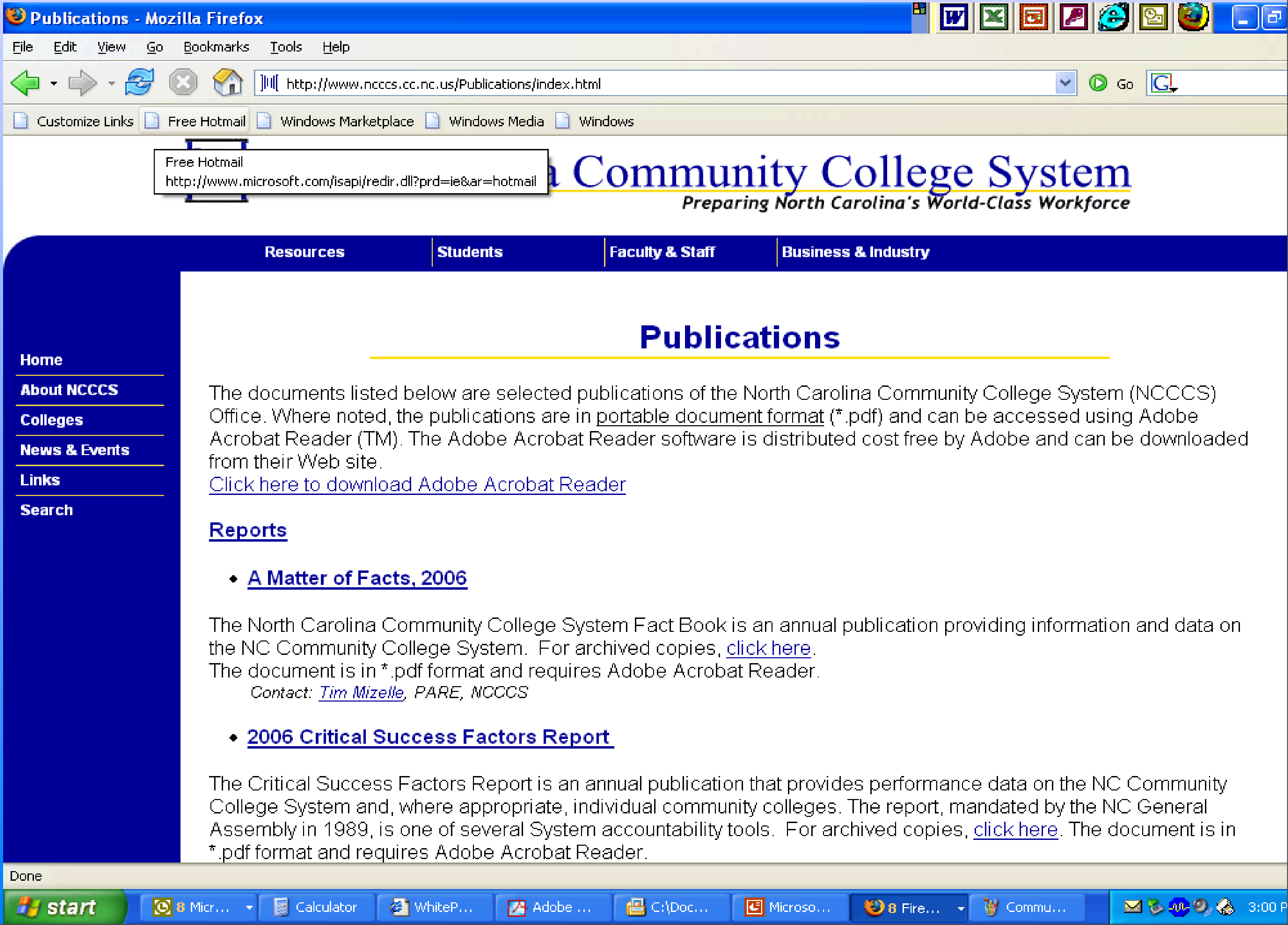
Catalog

Links

If you have comments or suggestions, about any page on our website, please note at the bottom of each page is the name of the person that is responsible for maintaining that page. Additionally, the name at the bottom of the page can be clicked on to send an email to that individual.

Or you may wish to direct your comments to the webmaster about technical issues related to management of the System Office web site, bug reports about forms pages, broken links, etc.

Additionally, you may prefer to search the NCCCS directory to locate a specific individual.



Publications

The documents listed below are selected publications of the North Carolina Community College System (NCCCS) Office. Where noted, the publications are in portable document format (*.pdf) and can be accessed using Adobe Acrobat Reader (TM). The Adobe Acrobat Reader software is distributed cost free by Adobe and can be downloaded from their Web site.
Click here to download Adobe Acrobat Reader

Reports

- A Matter of Facts, 2006

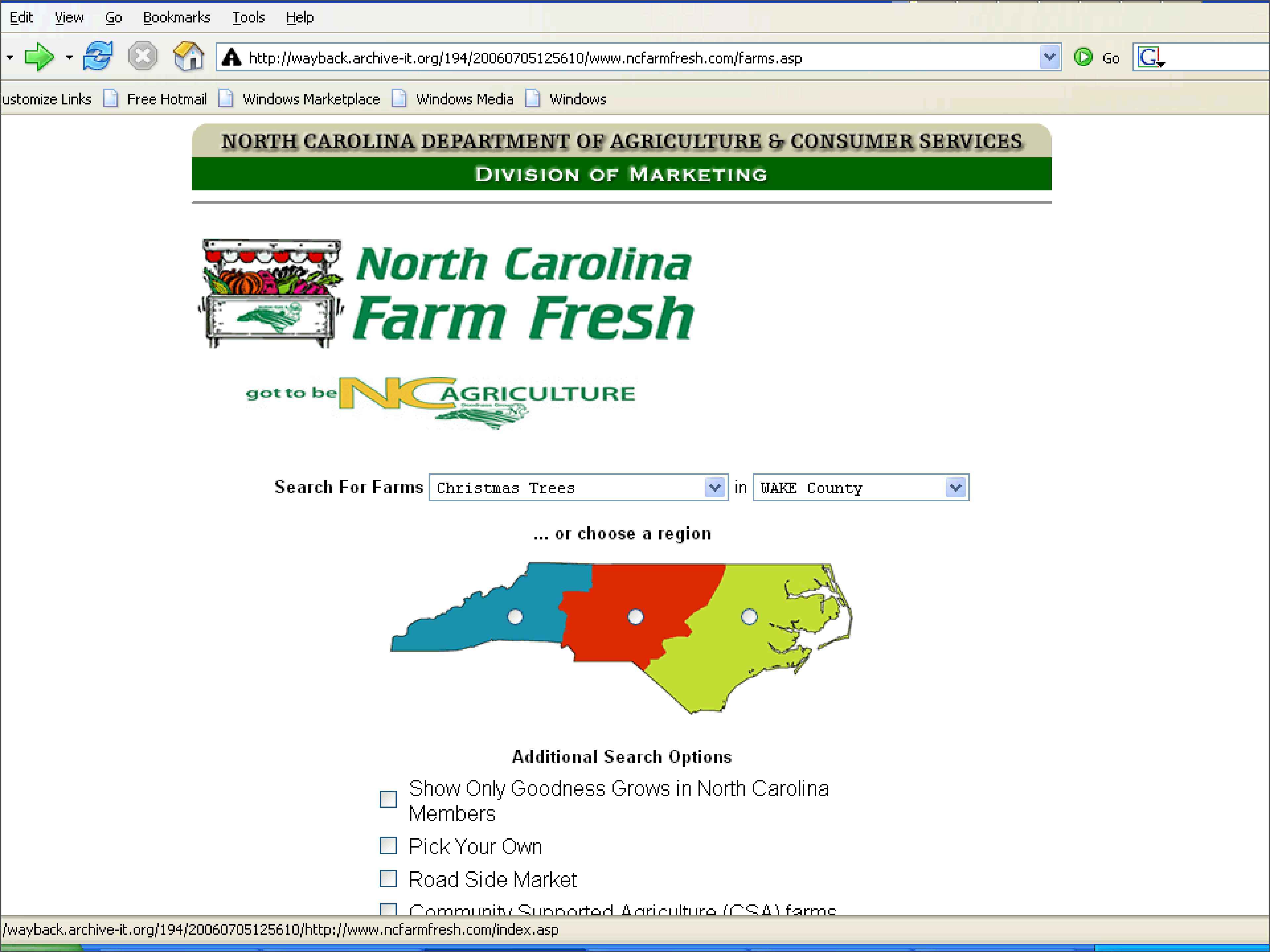
The North Carolina Community College System Fact Book is an annual publication providing information and data on the NC Community College System. For archived copies, click here.
The document is in *.pdf format and requires Adobe Acrobat Reader.
Contact: Tim Mizelle, PARE, NCCCS

- 2006 Critical Success Factors Report

The Critical Success Factors Report is an annual publication that provides performance data on the NC Community College System and, where appropriate, individual community colleges. The report, mandated by the NC General Assembly in 1989, is one of several System accountability tools. For archived copies, click here. The document is in *.pdf format and requires Adobe Acrobat Reader.

Discussion of Capture
What it Cannot Capture

- Archive-It and WAW cannot capture Streaming Video
- Archive-It and WAW cannot capture dynamic database driven sites
  - http://www.ncfarmfresh.com/
- Archive-It cannot capture password protected sites.



Edit View Go Bookmarks Tools Help

http://wayback.archive-it.org/194/20060705125610/www.ncfarmfresh.com/farms.asp

ustomize Links Free Hotmail Windows Marketplace Windows Media Windows

NORTH CAROLINA DEPARTMENT OF AGRICULTURE & CONSUMER SERVICES

DIVISION OF MARKETING

North Carolina Farm Fresh

got to be NC AGRICULTURE

Search For Farms Christmas Trees in WAKE County

... or choose a region

Additional Search Options

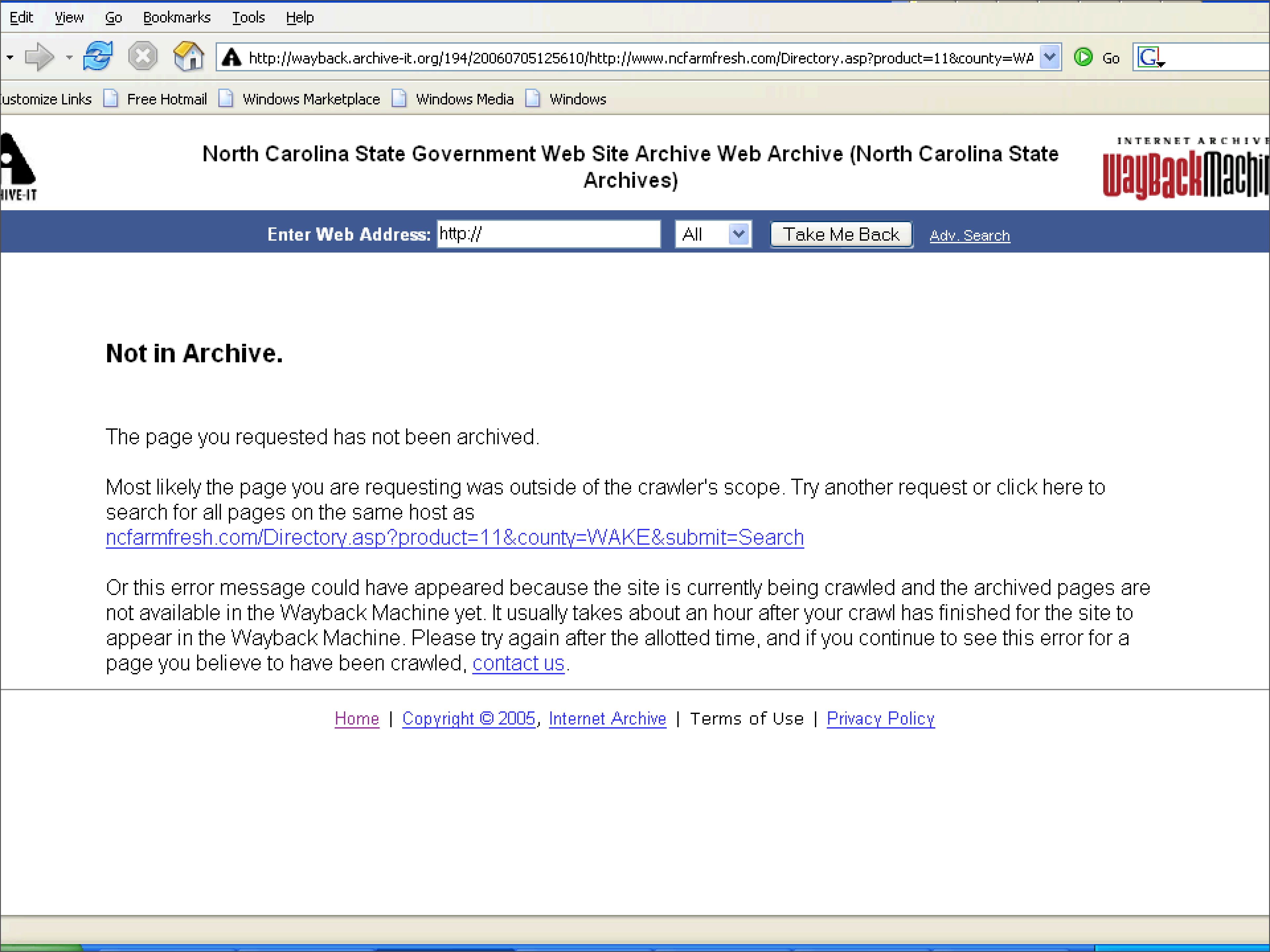
□ Show Only Goodness Grows in North Carolina Members

□ Pick Your Own

□ Road Side Market

□ Community Supported Agriculture (CSA) farms

/wayback.archive-it.org/194/20060705125610/http://www.ncfarmfresh.com/index.asp



North Carolina State Government Web Site Archive Web Archive (North Carolina State Archives)

Not in Archive.

The page you requested has not been archived.

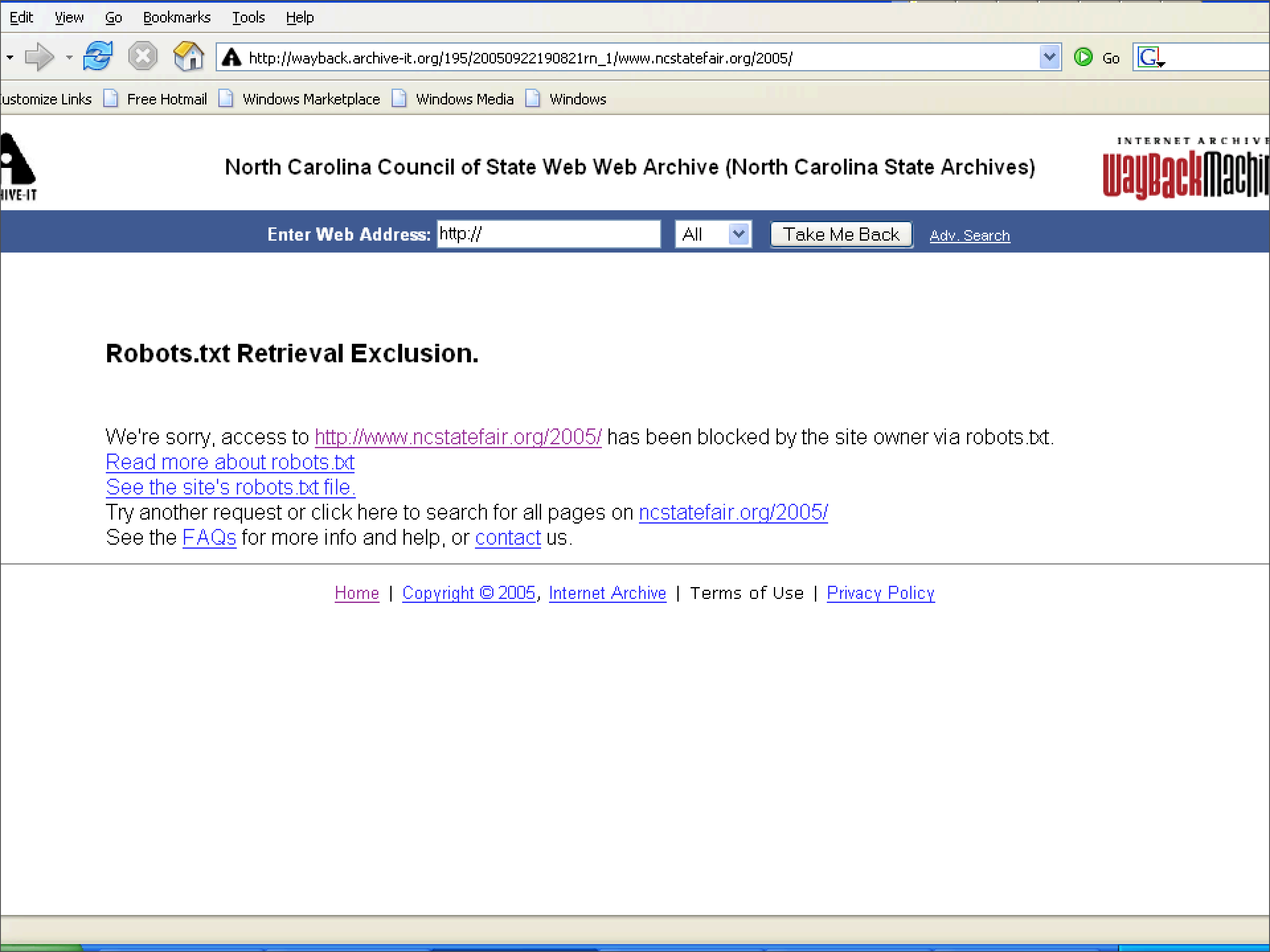
Most likely the page you are requesting was outside of the crawler's scope. Try another request or click here to search for all pages on the same host as ncfarmfresh.com/Directory.asp?product=11&county=WAKE&submit=Search

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.

Home | Copyright © 2005, Internet Archive | Terms of Use | Privacy Policy

Lessons Learned

- One hop off
  - “out of scope materials”
- Embedded Links Captured
  - “inappropriate materials”
    - Search for “lottery”
    - Embedded sponsor links (also in WAW)
  - Tremendous amount of material would have to be masked.
- Robots.txt
  - Example North Carolina State Fair.

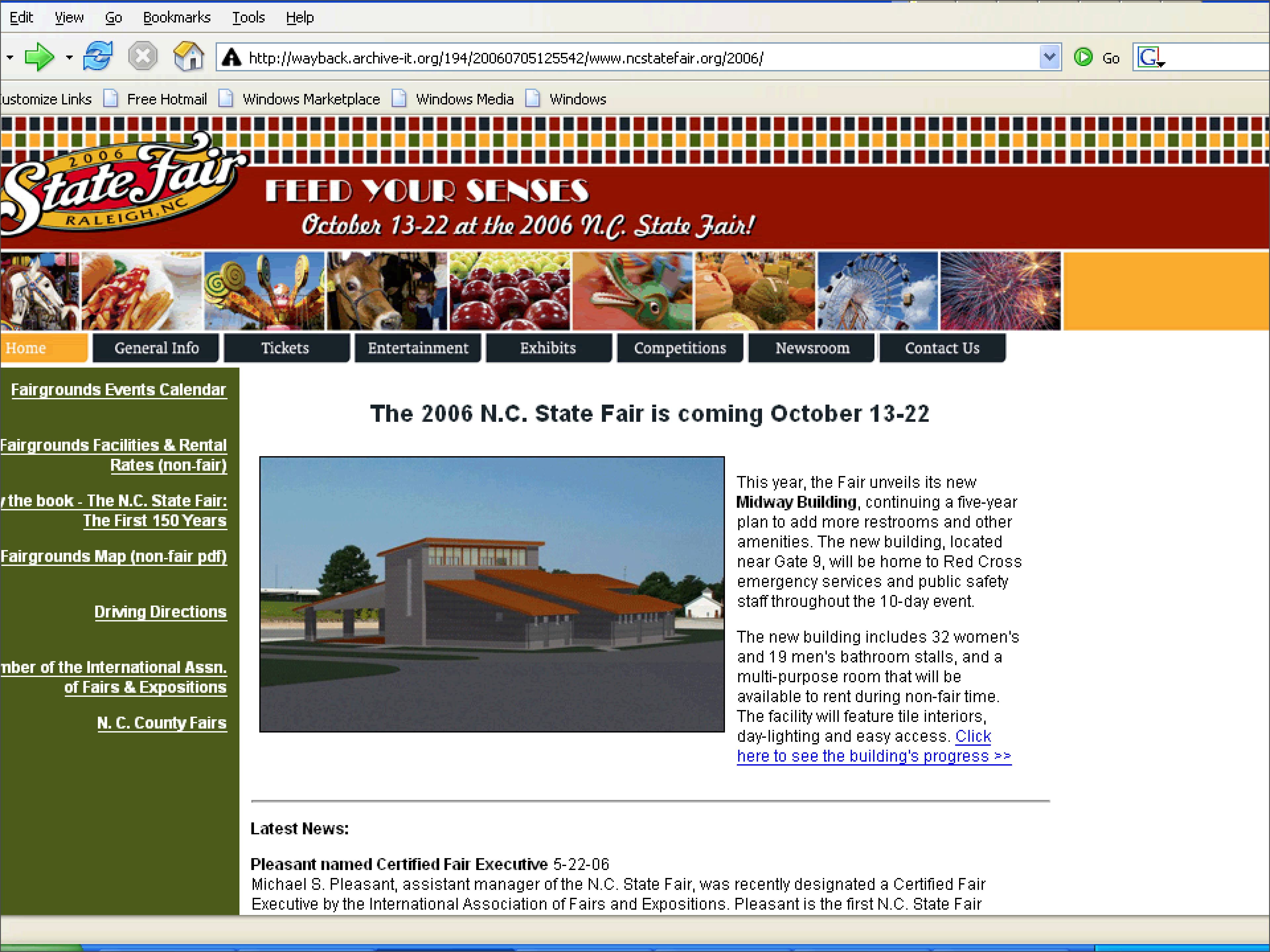

Edit View Go Bookmarks Tools Help
http://wayback.archive-it.org/195/20050922190821rn_1/www.ncstatefair.org/2005/
Go
ustomize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
IVE-IT
INTERNET ARCHIVE
WayBackMachi
Enter Web Address: http://
All
Take Me Back
Adv. Search

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

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/
See the FAQs for more info and help, or contact us.

Home | Copyright © 2005, Internet Archive | Terms of Use | Privacy Policy



http://wayback.archive-it.org/194/20060705125542/www.ncstatefair.org/2006/

FEED YOUR SENSES

October 13-22 at the 2006 N.C. State Fair!

Fairgrounds Events Calendar

Fairgrounds Facilities & Rental Rates (non-fair)

y the book - The N.C. State Fair: The First 150 Years

Fairgrounds Map (non-fair pdf)

Driving Directions

nber of the International Assn. of Fairs & Expositions

N. C. County Fairs

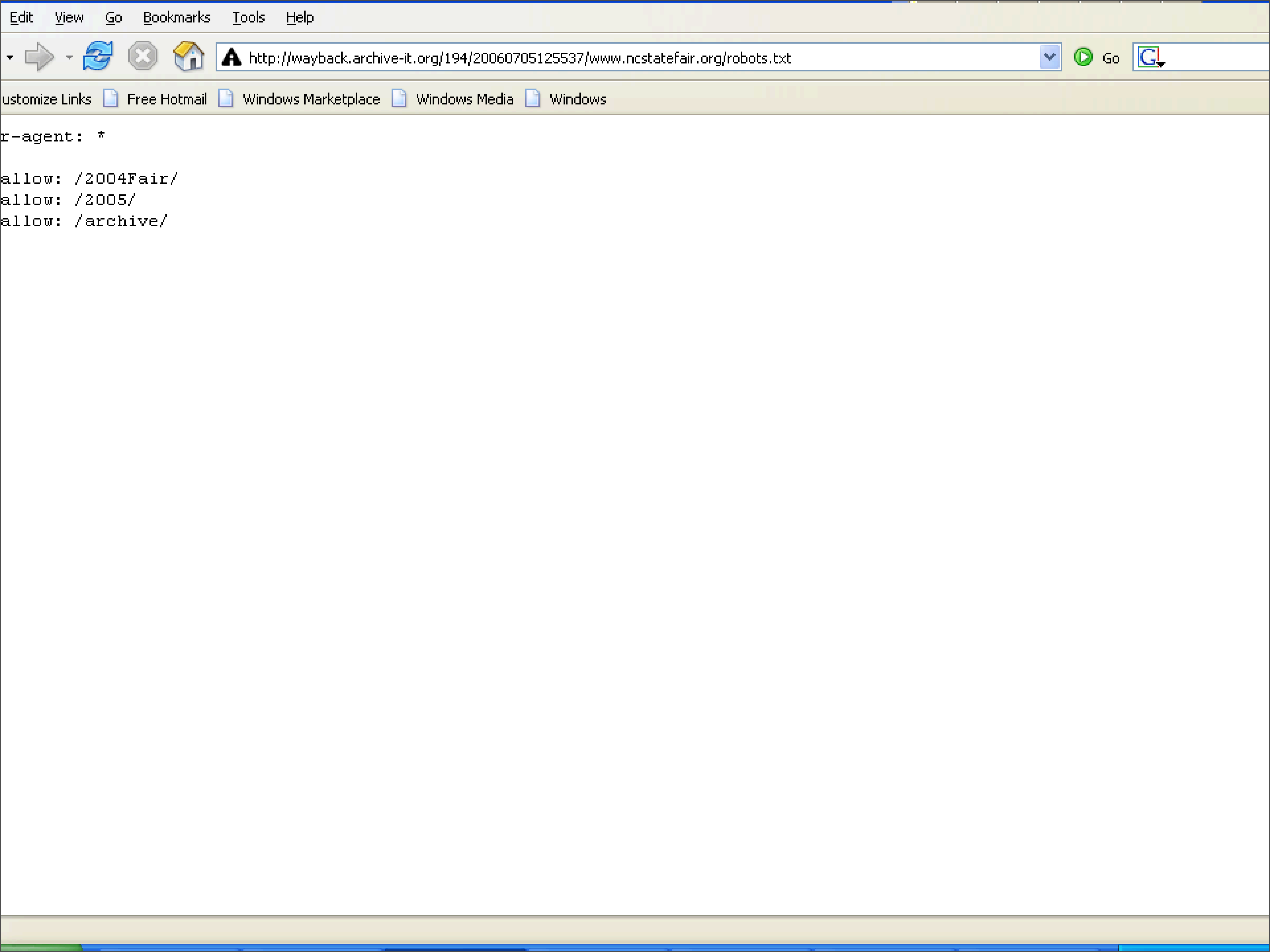
The 2006 N.C. State Fair is coming October 13-22

This year, the Fair unveils its new **Midway Building**, continuing a five-year plan to add more restrooms and other amenities. The new building, located near Gate 9, will be home to Red Cross emergency services and public safety staff throughout the 10-day event.

The new building includes 32 women's and 19 men's bathroom stalls, and a multi-purpose room that will be available to rent during non-fair time. The facility will feature tile interiors, day-lighting and easy access. Click here to see the building's progress >>

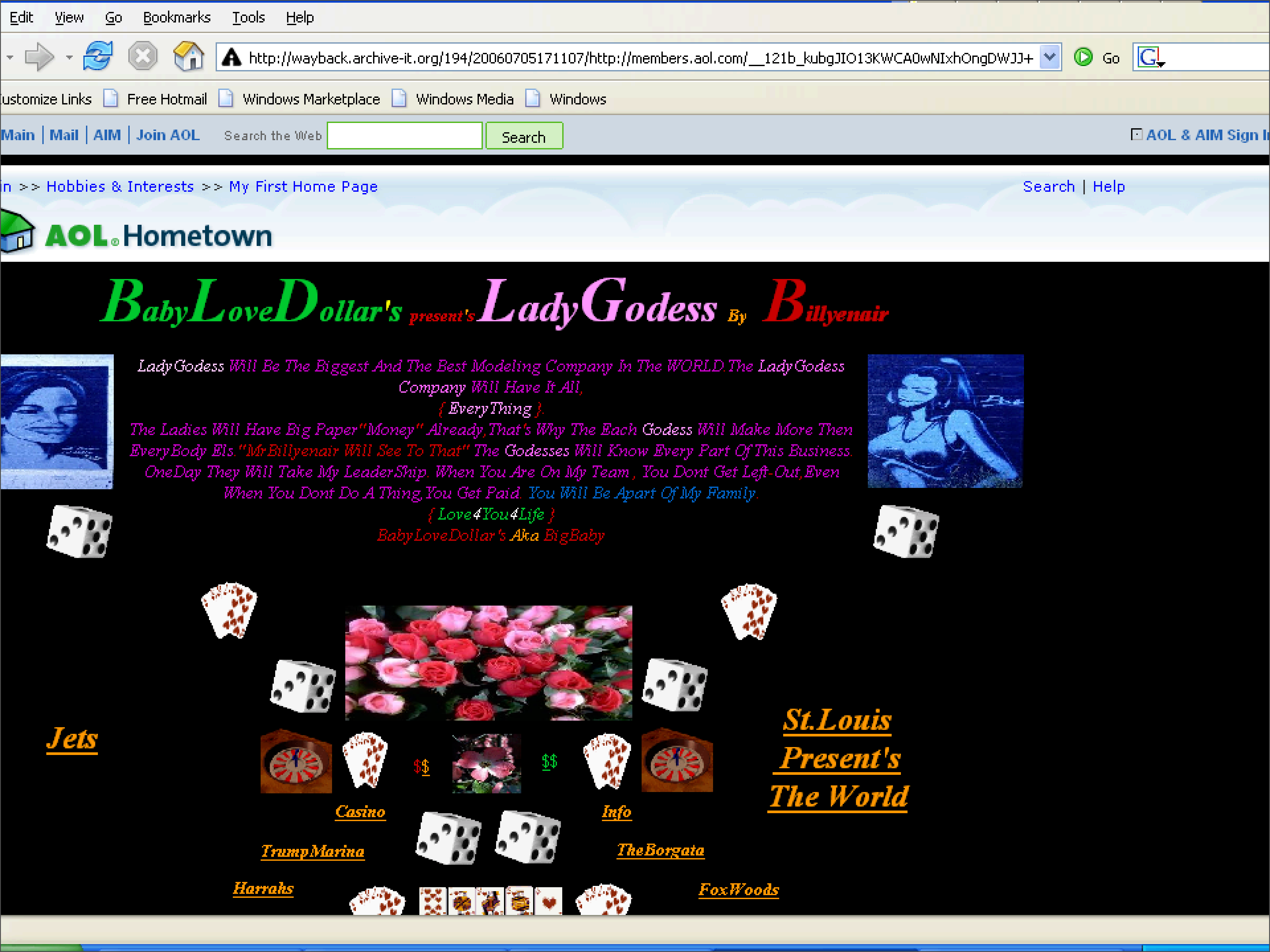
Latest News:

Pleasant named Certified Fair Executive 5-22-06
Michael S. Pleasant, assistant manager of the N.C. State Fair, was recently designated a Certified Fair Executive by the International Association of Fairs and Expositions. Pleasant is the first N.C. State Fair



```
r-agent: *

allow: /2004Fair/
allow: /2005/
allow: /archive/
```


Edit View Go Bookmarks Tools Help
http://wayback.archive-it.org/194/20060705171107/http://members.aol.com/__121b_kubgJIO13KWCA0wNIxhOngDWJJ+
Go
ustomize Links Free Hotmail Windows Marketplace Windows Media Windows
Main | Mail | AIM | Join AOL Search the Web Search
AOL & AIM Sign I
in >> Hobbies & Interests >> My First Home Page
Search | Help
AOL. Hometown

BabyLoveDollar's present's LadyGodess By Billyenair

LadyGodess Will Be The Biggest And The Best Modeling Company In The WORLD.The LadyGodess Company Will Have It All,
{ EveryThing }.
The Ladies Will Have Big Paper"Money" Already,That's Why The Each Godess Will Make More Then EveryBody Els."MrBillyenair Will See To That" The Godesses Will Know Every Part Of This Business. OneDay They Will Take My LeaderShip. When You Are On My Team , You Dont Get Left-Out,Even When You Dont Do A Thing,You Get Paid. You Will Be Apart Of My Family.
{ Love4You4Life }
BabyLoveDollar's Aka BigBaby

Jets

$$ $$

Casino

Info

TrumpMarina

TheBorgata

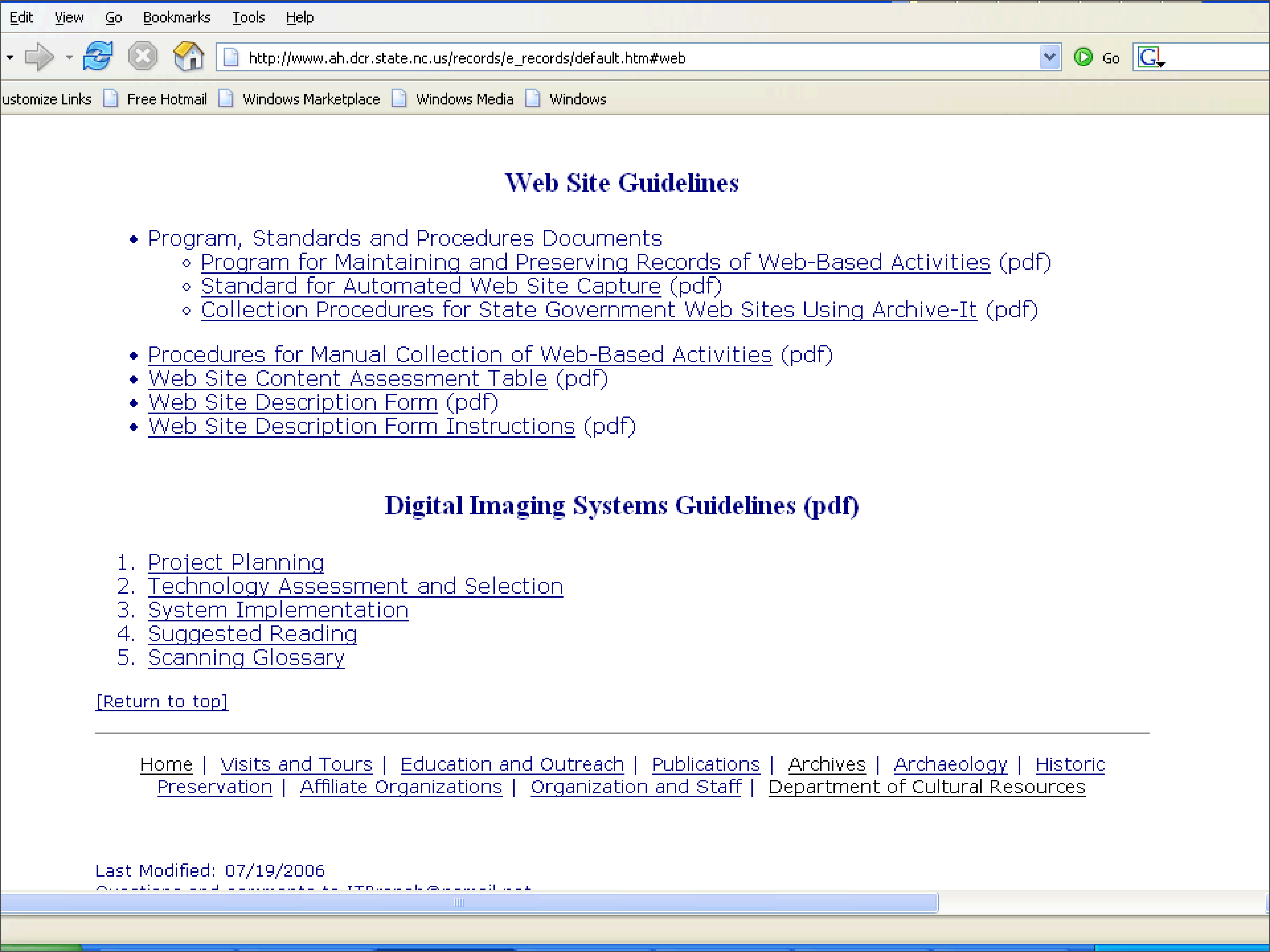
Harrahs

FoxWoods

St.Louis
Present's
The World

Analysis of 3 Methodologies

- **CEP**—on-site solution
  - One spider for each URL CVS format
  - Control over the crawler
  - Need institutional support
  - Need IT support.
  - Practitioner needs background in programming.
- **WAW**—hosted solution
  - Same crawler. Assume it will work similarly to Internet Archive.
  - Crawls specific web documents.
  - Control over the crawler
- **Internet Archive**—hosted solution
  - captures a moment in time
  - crawls everything within the domain.
  - Limited control over crawlers. Changed 7/2006.



Web Site Guidelines

- Program, Standards and Procedures Documents
  - Program for Maintaining and Preserving Records of Web-Based Activities (pdf)
  - Standard for Automated Web Site Capture (pdf)
  - Collection Procedures for State Government Web Sites Using Archive-It (pdf)

- Procedures for Manual Collection of Web-Based Activities (pdf)
- Web Site Content Assessment Table (pdf)
- Web Site Description Form (pdf)
- Web Site Description Form Instructions (pdf)

Digital Imaging Systems Guidelines (pdf)

1. Project Planning
2. Technology Assessment and Selection
3. System Implementation
4. Suggested Reading
5. Scanning Glossary

[Return to top]

Home | Visits and Tours | Education and Outreach | Publications | Archives | Archaeology | Historic Preservation | Affiliate Organizations | Organization and Staff | Department of Cultural Resources

Last Modified: 07/19/2006
Questions and comments to ITBranch@ncmail.net

Edit View Go Bookmarks Tools Help

http://www.ah.dcr.state.nc.us/archives/webarchive/index.html Go

ustomize Links Free Hotmail Windows Marketplace Windows Media Windows

North Carolina State Government Web Site Archive

North Carolina State Archives State Library of North Carolina

rch the Web Archive

Search

ch Help

Welcome!

nks

- me
- out Us
- owse by Agency
- owse by Collection
- e Map
- p
- ntact

her Sites

- chive-It
- ernet Archive

The North Carolina State Government Web Site Archive allows you to view North Carolina state agency web sites from past dates. The *Web Site Archive* 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 Archive* 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.

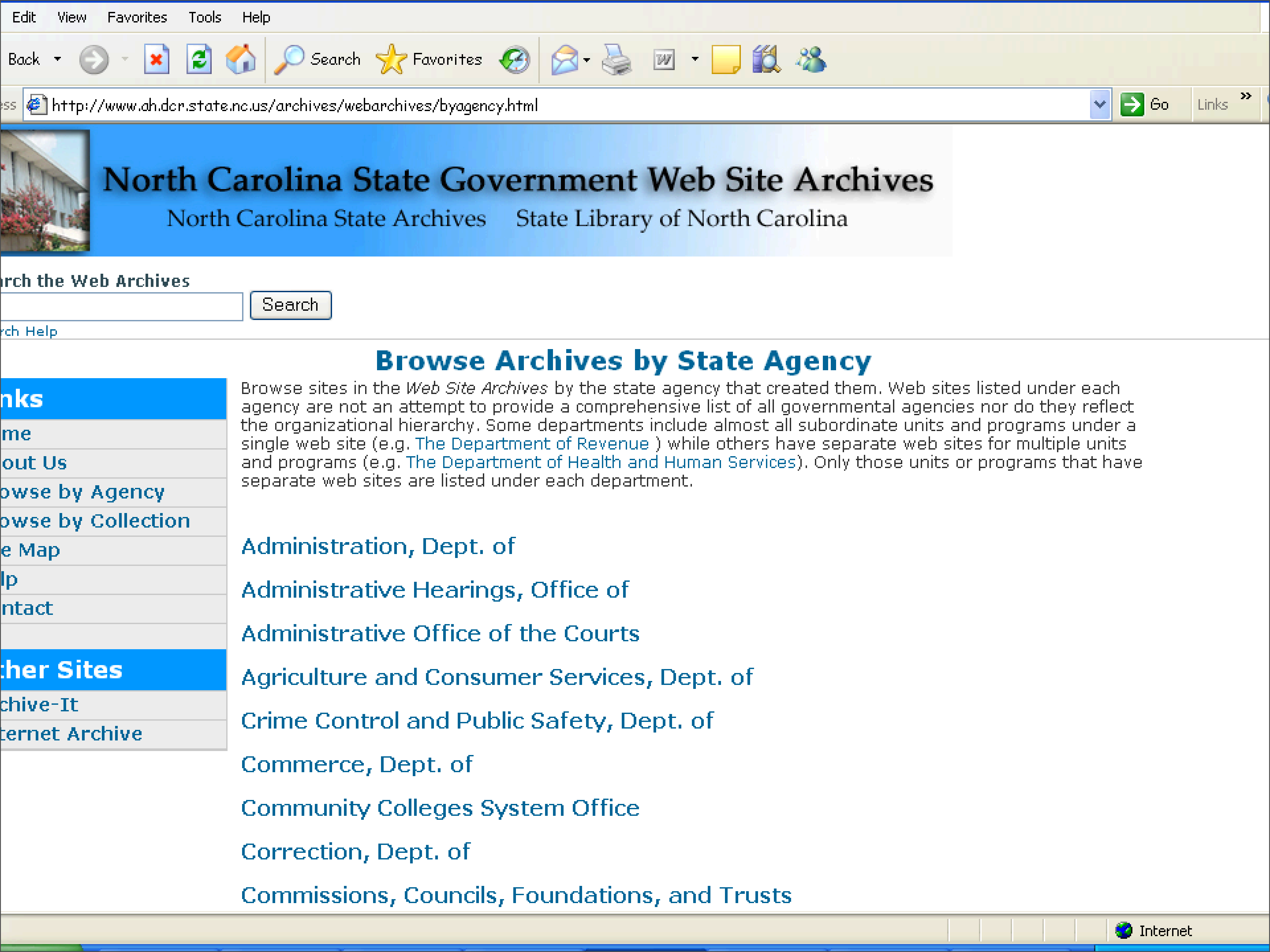
The Web Site Archive began as a pilot project with the Internet Archive during the Fall of 2005. The purpose of this project was to refine a tool called Archive-It which collects, preserves, and provides access to web sites of enduring value.

The success of the Pilot Project led to the creation of the current version of *The North Carolina State Government Web Site Archive* which began archiving web sites in late April of 2006. The *Web Site Archive* 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 Archive* 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.

Office of the Governor
STATE OF NORTH CAROLINA
HOME · MEET THE FIRST FAMILY · NEWS · CABINET INFORMATION · GOVERNOR'S OFFICE · CONTACT US
SEARCH THE SITE:
GOV. EASLEY WELCOMES KATRINA VICTIMS TO NC
GOV. EASLEY SIGNS EDUCATION LOTTERY BILL
Emergency Information
More News

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


Edit View Favorites Tools Help
Back
Search
Favorites
ess
http://www.ah.dcr.state.nc.us/archives/webarchives/byagency.html
Go
Links
Internet

North Carolina State Government Web Site Archives

North Carolina State Archives State Library of North Carolina

rch the Web Archives

Search

rch Help

Browse Archives by State Agency

nks

- me
- out Us
- owse by Agency
- owse by Collection
- e Map
- lp
- ntact

ther Sites

- chive-It
- ternet Archive

Browse sites in the *Web Site Archives* by the state agency that created them. Web sites listed under each agency are not an attempt to provide a comprehensive list of all governmental agencies nor do they reflect the organizational hierarchy. Some departments include almost all subordinate units and programs under a single web site (e.g. The Department of Revenue) while others have separate web sites for multiple units and programs (e.g. The Department of Health and Human Services). Only those units or programs that have separate web sites are listed under each department.

Administration, Dept. of

Administrative Hearings, Office of

Administrative Office of the Courts

Agriculture and Consumer Services, Dept. of

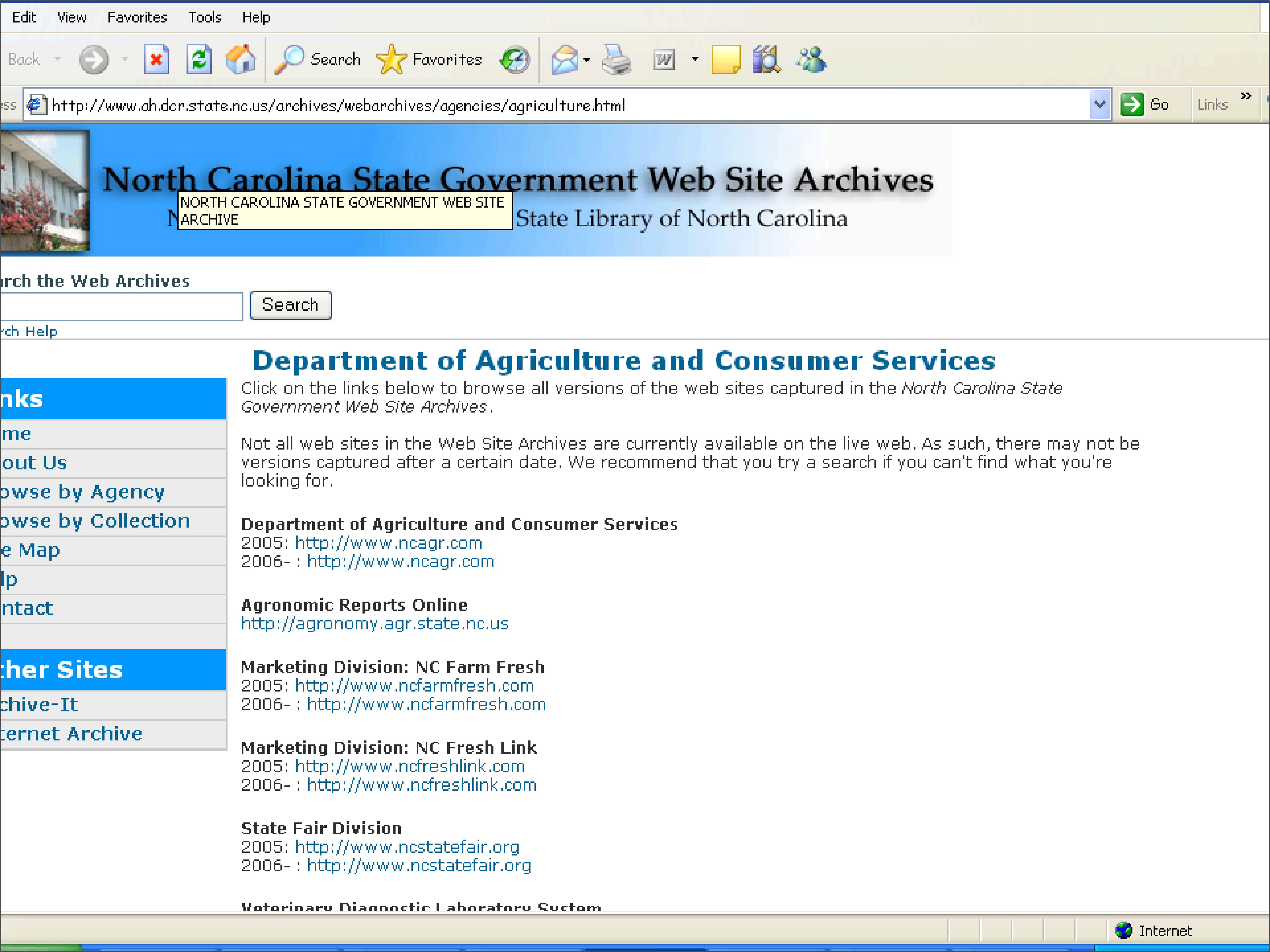
Crime Control and Public Safety, Dept. of

Commerce, Dept. of

Community Colleges System Office

Correction, Dept. of

Commissions, Councils, Foundations, and Trusts


Edit
View
Favorites
Tools
Help
Back
Search
Favorites
ess
http://www.ah.dcr.state.nc.us/archives/webarchives/agencies/agriculture.html
Go
Links
Internet

North Carolina State Government Web Site Archives

N NORTH CAROLINA STATE GOVERNMENT WEB SITE ARCHIVE State Library of North Carolina

rch the Web Archives

Search

rch Help

Department of Agriculture and Consumer Services

Click on the links below to browse all versions of the web sites captured in the *North Carolina State Government Web Site Archives*.

Not all web sites in the Web Site Archives are currently available on the live web. As such, there may not be versions captured after a certain date. We recommend that you try a search if you can't find what you're looking for.

Department of Agriculture and Consumer Services
2005: http://www.ncagr.com
2006- : http://www.ncagr.com

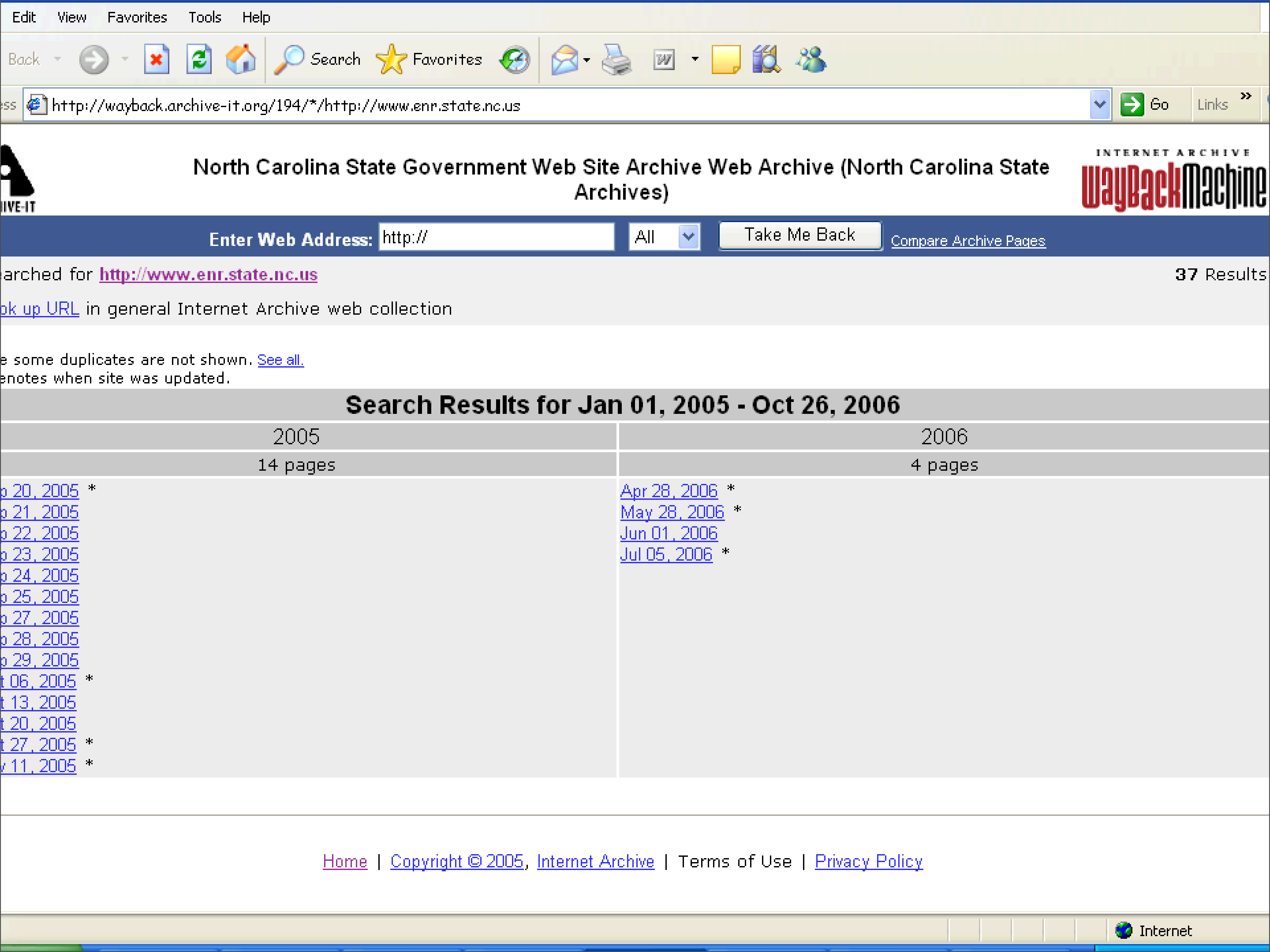
Agronomic Reports Online
http://agronomy.agr.state.nc.us

Marketing Division: NC Farm Fresh
2005: http://www.ncfarmfresh.com
2006- : http://www.ncfarmfresh.com

Marketing Division: NC Fresh Link
2005: http://www.ncfreshlink.com
2006- : http://www.ncfreshlink.com

State Fair Division
2005: http://www.ncstatefair.org
2006- : http://www.ncstatefair.org

Veterinary Diagnostic Laboratory System



http://wayback.archive-it.org/194/*/http://www.enr.state.nc.us

North Carolina State Government Web Site Archive Web Archive (North Carolina State Archives)

Enter Web Address: http:// All Take Me Back Compare Archive Pages

arched for http://www.enr.state.nc.us 37 Results

ok up URL in general Internet Archive web collection

e some duplicates are not shown. See all.
enotes when site was updated.

Search Results for Jan 01, 2005 - Oct 26, 2006

2005	2006
14 pages	4 pages
p 20, 2005 *	Apr 28, 2006 *
p 21, 2005	May 28, 2006 *
p 22, 2005	Jun 01, 2006
p 23, 2005	Jul 05, 2006 *
p 24, 2005	
p 25, 2005	
p 27, 2005	
p 28, 2005	
p 29, 2005	
t 06, 2005 *	
t 13, 2005	
t 20, 2005	
t 27, 2005 *	
v 11, 2005 *	

Home | Copyright © 2005, Internet Archive | Terms of Use | Privacy Policy

Contact Information

Kelly Eubank
Electronic Records Archivist
North Carolina Archives and History
Telephone: (919) 807-7355
Email: kelly.eubank@ncmail.net
Web:
http://www.ah.dcr.state.nc.us/archives/webarchive/index.html
http://www.ah.dcr.state.nc.us/records/default.htm

Jennifer Ricker

State Library of North Carolina

(919) 807-7455

Email: jricker@library.dcr.state.nc.us

Web address: http://statelibrary.dcr.state.nc.us/dimp/index.html