North Carolina Department of Cultural Resources
Office of Archives and History

Procedures for Manual Collection 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).

DCR has initiated a new comprehensive program to handle important information now available through agency Web sites (see *Program for Maintaining and Preserving Records of Web-Based Activities* for the legal basis of this program). Beginning in September, 2005, DCR began collecting Web sites via an automatic crawling service to become part of its permanent collection. DCR uses the tool Archive-It, developed by the Internet Archive, to collect, store, and provide access to these Web sites. The technological capability of Archive-It guides these procedures. The Archive-It software cannot capture Web-enabled databases which require user input, and sites with heavy JavaScript may also be precluded from capture (See *Standard for Automated Web Site Capture* and *Collection Procedures for State Government Web sites using Archive-It* for more information about the Archive-It tool and frequency of capture).

Participation in automated capture for agency Web sites is not mandatory, but DCR strongly recommends it for ease of capture on both the part of DCR and the part of the agency. However, because Web sites are legal records under G.S. 132, agencies who prefer not to participate in automated capture, or who cannot participate because of technical considerations, are not exempt from providing DCR with a copy of their Web site(s). Agencies may instead manually provide DCR with copies of their Web site(s). Web-enabled databases, while not captured automatically through Archive-It, present a different set of challenges and need to be collected through other means yet to be determined.

What to Include in the Web Site Snapshot

If an agency elects not to participate in the automatic crawling performed by Archive-It, or their Web site(s) cannot be crawled due to JavaScript limitations, DCR recommends that the agency take a snapshot of their Web site(s). DCR recommends that Web site snapshots be taken at the time of each major version change to the Web site (different look, additional features, etc.) or at least once a year, whichever occurs first. For those agencies having a high litigation risk, DCR recommends that the agency audit every change to the site, cite the date that change occurred and whether or not that change was posted as part of the official Web site. This documents for legal purposes the agency's position at a particular point in time.

Agencies should include all active documents available to the public that are located on the agency's Web server, including copies of agency documents that exist in another form elsewhere, EXCEPT:

1. Databases
2. Files located on a Web server external to the agency (e.g. another agency's Web site).

Web site Description Form

In order to properly preserve a Web site, DCR requests that agencies also complete a standard description form (see attachment) developed by DCR. This form permits government agencies to capture easily information about the content, format, and technical characteristics of their Web sites. Submission of this descriptive information, along with copies of all active source files and both electronic and hard copy versions of relevant log files that document the names of files supporting the Web site, will allow DCR to provide continuing access to and an historical perspective for the provision of government information and services. (Agencies may deem it efficient to capture relevant source files using Web site capture software. This is an acceptable alternative to identifying and capturing source files manually, as long as the files are saved as ASCII text.)

Acceptable Media for Submission of Web site Snapshots

To ensure ease of transfer and standardization of the media being submitted for preservation, DCR suggests that agencies adhere to the following media specifics when capturing their Web sites:

1. Use a fresh CD-R.
2. A gold reflective surface is preferred but not required.
3. Source files should not be compressed.
4. The CD's case should be appropriately labeled, but the CD itself should not. A volume label is automatically created during initial use of the CD, and may be edited to reflect your agency's nickname/acronym, date, and/or disk number within the allotted 11 characters. Place this number on the case and all related transfer documentation.
5. The CD writing should comply with the Joliet modifications to ISO 9660 specifications. These settings are available in your CD creation software options.

Procedure for Submission of Web site Snapshots to NC Office of Archives and History

The media and documentation from the Web site snapshot may be transferred to the Office of Archives and History using the standard records transfer form (see http://www.ah.dcr.state.nc.us/sections/archives/rec/transfer.htm or call the State Records Center at 919-807-7350). The media and documentation may only be transferred under provisions of an approved records retention and disposition schedule and will be transferred immediately to the custody of the North Carolina State Archives upon receipt.

Attachments

Web site Content Assessment Table
http://www.ah.dcr.state.nc.us/records/e_records/default.htm#web
Web site Description Form
http://www.ah.dcr.state.nc.us/records/e_records/default.htm#web
Web site Description Form Instructions
http://www.ah.dcr.state.nc.us/records/e_records/default.htm#web

Definitions

Crawl – the information captured by a crawler on a single visit to each of the specified urls.

Crawler – a software agent that captures information from the Web. Archive-It's crawler starts with a list of urls to visit. As it visits these urls it captures the documents on these Web pages.

HTML – a non-proprietary file format for describing the structure of hypermedia documents—plain text (ASCII) files with embedded codes for logical markup, using tags to structure text into table, interactive forms, headings, paragraphs, lists, and more. It can be created and processed with a wide range of tools from simple text editors to sophisticated authoring software.

Web – a decentralized, global network connecting millions of computers. It allows computer users to communicate information to each other.

Web page – a source file, maintained as ASCII text, provided by a file server, and subsequently executed on a local computer, that uses HTML/XML markup languages and external software tools to produce a representation and provide meaning.

Web sites – collections of information, documents, and databases that are provided to a user community utilizing World Wide Web formats and protocols.

Web site Content Assessment

To help with the analysis of an agency's Web site, the Office of Archives and History has included a Web site Content Assessment (WCA) Table, which will allow an agency to determine the level of legal and/or managerial risk associated with their Web site and their online presence. An agency's future records keeping actions will be determined by the risk level under which they are currently operating.