North Carolina Department of Cultural Resources

Collection Procedures for State Government Web sites using Archive-It

As the Internet Archive releases more features in its Archive-It tool, this document will be revised to reflect those changes.

The North Carolina Department of Cultural Resources (DCR) collects Web sites 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:

- Appraisal decisions must be made at the domain name level (called seeds in Archive-It).
- There can be up to 3 collections with up to 100 seeds per collection or 1 large collection with 300 active seeds.
- Each seed can be captured according to a different time interval.
- There is a limit of 10 million unique objects collected per year, up to .5 terabytes of data.
- The software cannot capture pages generated from web-enabled databases, which require user input.
- The software can capture information written in JavaScript but cannot always render that information in Archive-It's viewer, the Wayback Machine. In this case the researcher is sent out to the live Web.

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 capture based on content, DCR will investigate alternative methods for capturing the information.
- Web sites that rely heavily on scripting for display may be excluded by the crawler because of technical limitations preventing access to the captured information. Current technologies cannot capture this type of information.

Application of the Web Site Macro-Appraisal Score

All Web sites defined as “In Scope” and able to be captured will be assigned a Web site Macro-Appraisal Score using the point system described in the Web site Capture Standards document. The cut-off point for inclusion into the Web archive will depend both on the value of the Web site and on the size limitations of the Archive-It service. This score will also determine the capture frequency. The seed inclusion list may be adjusted throughout the year as new Web sites are discovered, and as Web sites become defunct. Crawl frequency also may be adjusted depending on the size and growth rate of the archives. If a domain falls out of scope due technological limitations, the site will not be crawled. The agency needs to work with the DCR staff to ensure that the site is preserved. DCR staff utilizes reports provided by Internet Archive to identify those sites that the crawler cannot capture and will contact staff at the state agency to begin discussions regarding alternative means of capture and preservation.

Frequency of crawl, based on Macro-Appraisal Score (4/2806-6/15/06):

7.0-7.99:	Crawled annually
8.0-10.99:	Crawled quarterly
11-21:	Crawled monthly

Frequency of crawl, based on Macro-Appraisal Score (6/15/06-)

7.0-7.99:	Crawled annually
8.0-21:	Crawled quarterly

Web Archive Analysis and Management

The Internet Archive has agreed to submit to its users a detailed report listing the captured domains, the byte size of the captured documents, the number of documents captured as well as a breakdown of the file formats collected. DCR staff analyzes these reports to assess the rate of growth and the types of information captured. If the growth rate of a particular domain increases rapidly, the staff will adjust the crawl frequency in order to stay within the parameters set forth in the scope of services. Additionally, if the staff discovers that a particular domain generates a tremendous amount of material that is out of scope, it may exclude or delete the domain from its crawl in order to conserve space limitations and/or limit liability.