State Government Web Site Collection Guideline

The Virginia Public Records Act of 2006 (§42.1-76 et seq.) amends, repeals, and reenacts several sections of the Code of Virginia and provides legislation to facilitate the acquisition, maintenance, and preservation of Virginia public records, including materials in electronic format.

It will be the practice of the Library of Virginia to collect, preserve, and provide access to all Web sites of Virginia's state government agencies in the executive, legislative, and judicial branches of government as described in this guideline. State government Web sites will be selected for inclusion based on intellectual content, research and educational use, and long-term benefit to the citizens of the commonwealth.

A state government Web site is defined as the collection of all files identified by a state government domain for the purpose of providing publicly available information, affording access to government services, and/or conducting the state's business. Large, complex Web sites may span several servers and potentially several domains but are unified by Virginia government related content.

All Web sites selected will be collected and preserved in the formats in which they were primarily distributed to the public. They will be made accessible from the Library of Virginia's catalog and/or Web site, as well as the Internet Archive's Archive-It Web site (http://www.archive-it.org).

The Library of Virginia will collect the following Web sites:

- All executive, legislative, and judicial branch state agencies, commissions and boards as listed in *The Report of Secretary of the Commonwealth* (i.e. *Bluebook*)
- All independent state agencies listed in *The Report of the Secretary of the Commonwealth* (i.e. *Bluebook*)
- All statewide constitutional officers (e.g. Office of the Governor, Office of the Lt. Governor, and Office of the Attorney General), the Governor's Cabinet Secretaries, and the First Lady
- Gubernatorial initiatives and special projects (e.g. Smart Beginnings, Capitol Square Renovations, Springfield Interchange, Jamestown 2007)

Web sites are collected on an established schedule, currently monthly for statewide constitutional officers, the Governor's Cabinet, Gubernatorial initiatives, and the First Lady, while all remaining Web sites will be crawled quarterly. In the occurrence of significant public safety and health incidents or other noteworthy events, the Library of Virginia may, at its discretion, alter the crawling frequency of state agency Web sites. Due to the dynamic nature of Web archiving technology, this guideline will be reviewed and revised regularly.

We will not collect:

- Public or private college and university Web sites. Due to the extent of these Web resources, access restrictions, and existing state Archives practice these sites will not be archived.
- Non-state government Web sites. In general, captured Web sites will contain only state government information. Non-state government Web sites may be considered for capture if they contain significant state government information and assist in the formation of government policy.

Technical limitations:

The Library of Virginia has partnered with the Internet Archive to collect, preserve and provide access to the Library's Web archive collections to the best of our abilities via the Archive-It service. Web content is harvested using the Heretrix Web crawler and archived content is delivered to users via the Wayback Machine.

As a general rule of thumb, simple static web pages are the easiest to archive. Limitations to capturing and playing back archival Web content are as follows:

- When a dynamic page contains forms, JavaScript, images, streaming media or other elements that require interaction with the originating host, the archived pages may not contain the original site's functionality.
- Database driven Web sites can be very difficult to harvest. For example, if you need to fill in a form to get access to the content, such as with a search box, the harvester typically cannot retrieve the content.
- JavaScript elements are often hard to archive and even harder to display in the Wayback Machine, especially if they generate relative links (links which do not contain the full address of the linked page).
- Web site owners can specify files or directories that are disallowed from a crawl, and they can even create specific rules for different automated crawlers. All of this information is contained in a file called **robots.txt**. The Archive-It tool respects robots.txt exclusion headers. The Library will make every effort to contact site owners to be sure that they allow the Archive-It crawler to have appropriate access to their site.
- Password protected sites cannot be accessed by the crawler and therefore will not be archived.
- Links to sites that are not in the same domain as a url identified for archiving will not be captured. For example, if the Secretary of Public Safety site has a link to the Red Cross (www.redcross.org), the Red Cross's site will not be captured. However embedded files on seed pages are crawled regardless of whether they come from an offsite host or not.