Political and Organizational Web Site Collection Development Guideline

The Library of Virginia collects, preserves, and provides access to all Web sites of Virginia's state government agencies in the executive, legislative, and judicial branches of government. State government Web sites are selected for inclusion based on intellectual content, research and educational use, and long-term benefit to the citizens of the commonwealth. See ***State Government Web Site Collection Guideline*** for details.

While these guidelines ensure the capture of official state government Web sites, it does not include many other important non-governmental Web sites that help provide a fuller context in which political and governmental decisions are made in Virginia. The Library of Virginia will collect, preserve and provide access to the Web sites of individual members of the Virginia legislative branch, Congressional delegation, candidate sites, political parties and other sites as described in this guideline.

In addition, the Library of Virginia will collect, preserve and provide access to the Web sites of organizations that are already represented in the Library's Private Papers collection. These will include but are not limited to social, civic, cultural, service and other groups that may be either entirely Virginia-based, or the Virginia branches of larger organizations.

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 from the Archive-It Web site (http://www.archive-it.org), a subscription service of the Internet Archive, a non-profit organization founded to build an Internet library offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format.

The Library of Virginia will collect the following Web sites for:

- Web sites of individual members of the Virginia legislative branch, Congressional delegation, candidate sites, political parties, Political Action Committees, "Sunshine" organizations, and other sites at the discretion of project archivists.
- Web sites of Virginia organizations (state and local level) for which the Library of Virginia has archival collections.
- Web sites of special events that bring national attention to the state, for example, The Richmond Folk Festival and Virginia Wine Festival.
- Other sites at the discretion of the project archivists

Web sites are collected on an established schedule (to-be-determined). The Library of Virginia may, at its discretion, alter the crawling frequency. Due to the dynamic nature of Web archiving technology, this guideline will be reviewed and revised regularly.

Permissions:

The Library of Virginia will send each non-governmental Web site a permission request asking to crawl their site. The permission request will list the specific url to be crawled, describe the collection it will be a part of and contain a link to an on-line permission form. If the Library does not receive a response to the initial request, staff members will send a second request. If staff does not receive a response after two requests, the collections will be crawled under the "fair use" provision of the Copyright Act.

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 its abilities via the Archive-It service. Web content is harvested using the Heritrix Web crawler and archived content is indexed and searchable via the Internet Archive's Wayback Machine.

As a general rule, 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 might 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 often are hard to archive and even harder to display in the Wayback Machine, especially if they generate relative links (links that do not contain the full address of the linked page).
- Web site owners can specify files or directories to be excluded from a crawl, and 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 are crawled regardless of whether or not they come from an offsite host.