Tools and Techniques for Harvesting the World Wide Web

Jennifer L. Marill, Andrew Boyko, Michael Ashenfelder, and Laura Graham
Office of Strategic Initiatives
Library of Congress
101 Independence Avenue, SE
Washington, DC 20540
{jmarill, aboy, mashe, lgrah}@loc.gov

Categories and Subject Descriptors: H.3.7 [**Digital Libraries**]
General Terms: Experimentation.

Keywords: Web Harvesting, Web Archiving, Digital Preservation, Harvesting Tools.

ABSTRACT
Recently the Library of Congress began developing a strategy for the preservation of digital content. Efforts have focused on the need to select, harvest, describe, access and preserve Web resources. This poster focuses on the Library's initial investigation and evaluation of Web harvesting software tools.

WEB HARVESTING TOOLS
While there are a number of tools for Web crawling, only a small subset of tools can handle the more specific and challenging tasks of large-scale harvesting for long-term preservation. From this subset, the Library is examining the following tools based on these criteria: open source, documented prior use, and an active community of developers and users.

Our minimum test platform is a 1.8GHz Pentium 4, with 768 MB of RAM, and hundreds of gigabytes of disk space. We run Fedora Linux because of its mainstream usage and active community support. We are benchmarking the various processes in an attempt to estimate their scalability.

HTTrack [1], a desktop crawler, is easy to configure, widely used and available for both Windows and Unix systems. However, HTTrack is best suited for an exploratory acquisition of a small number of sites. It modifies the links in retrieved content to create a self-consistent set of files that can be directly viewed without the need of a separate viewing tool. HTTrack is valuable for site analysis but not suitable for wide-scale harvesting.

The open source NEDLIB Harvester [2], developed by a European consortium, is used by a number of national libraries. Despite its popularity in this small community, its development has been dormant since September 2002. NEDLIB relies on a relational database (MySQL) for its configuration and process control. While the database adds complexity to its use, it also provides the ability for extensive reporting. In our initial testing, NEDLIB's crawl configuration - performed by adding values for seeds, inclusions, and exclusions to database tables - was sufficiently expressive for general crawls. However, NEDLIB does not have the flexibility required by more complex permissions environments. The harvester lacks a direct user interface and communicates its progress through logging and database entries. Integrating NEDLIB into a regular harvesting workflow will require the development of a superstructure of tools. We are still establishing approaches for measuring crawler performance and quality, but were satisfied with NEDLIB's results on moderately sized (tens of gigabytes), narrowly scoped crawls.

Heritrix [3], initiated by the Internet Archive [4], was released to the public as recently as January 2004. The Library's initial testing of Heritrix has been promising, and the public nature of the software's development process instills confidence in its future improvement. Heritrix is driven by an XML configuration language, which supports complex crawl definitions and filtering. In addition, it appears to support advanced customization via Java plug-ins. Heritrix includes a Web hosted control panel for managing and monitoring crawls. Based on the quality of our results and our positive impression of the development process to date, we will continue to investigate Heritrix for a variety of crawls.

ACCESS TOOLS
Both Heritrix and NEDLIB require a separate tool to handle the presentation of harvested content. The Internet Archive's Wayback Machine [5], a proprietary presentation tool used for Alexa Internet [6] and Heritrix output, is difficult to configure for in-house use. The NWA Toolset [7] is an open source presentation tool designed to display the output of the NEDLIB harvester; however, the toolset's modular design makes it compatible with Heritrix output. The NWA Toolset's first published release (December 2003) was challenging to configure and deploy, and is demanding on hardware resources. Nevertheless, we have not found any major architectural problems with it, and continue to actively test it.

REFERENCES
[1] HTTrack. http://www.httrack.com

[2] NEDLIB Harvester. http://www.csc.fi/sovellus/nedlib/

[3] Heritrix. http://crawler.archive.org/

[4] The Internet Archive. http://www.archive.org/

[5] The Internet Archive Wayback Machine. http://www.archive.org/about/faqs.php#The_Wayback_Machine

[6] Alexa Internet. http://pages.alexa.com/prod_serv/alexa_crawl.html

[7] NWA Toolset. http://nwa.nb.no/index.php?doc=download

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.
ACM 1-58113-832-6/04/0006.

Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)
1-58113-832-6/04 $ 20.00 © 2004 ACM