Getting the Tar Off Our Heels: Moving Forward with Archiving University of North Carolina at Chapel Hill Websites

Lisa Gregory, Carolina Digital Curation Fellow
School of Information and Library Science, University of North Carolina at Chapel Hill

Setting

Fellowship Description

The University Archives and Records Management Services (UARMS) at the University of North Carolina at Chapel Hill is exploring the possibility of archiving websites as a continuation of its current collecting mandate. During the 2008-2009 school year, the author, a DigCCurr fellow assigned to UARMS, investigated the feasibility of integrating website archiving into the archives' workflow.

Fellowship Goals

- Identify open source tools for archiving websites.
- Find documentation of case studies involving identified tools.
- Install tools that seem most viable for UARMS.
- Run test harvests and explore features of installed tools.
- Recommend tools, policies, and procedures for UARMS use.

Tool Overview

NetarchiveSuite (Quick Start Version)

Creators: Danish Royal Library and State and University Library

Purpose: "plan, schedule and run web harvests of parts of the Internet. It scales to a wide range of tasks, from small, thematic harvests . . . to harvesting and archiving the content of an entire national domain." (http://netarchive.dk/suite)

Functions included: harvesting, quality assurance and review, bit preservation

HTTrack

Creators: Xavier Roche with other contributors

Purpose: an offline browser utility

Functions included: mirroring of site, conversion to relative link structure

Web Curator Tool (WCT)

Creators: National Library of New Zealand and British Library

Purpose: "a tool for managing the selective web harvesting process. It is designed for use in libraries and other collecting organisations, and supports collection by non-technical users while still allowing complete control of the web harvesting process." (http://webcurator.sourceforge.net)

Functions included: tracking harvest permissions, harvesting, adding basic Dublin Core metadata, quality assurance and review

Method and Criteria

Tools were evaluated using a rubric designed to reflect the requirements and desires of UARMS. The four categories and their related criteria are listed below. **Ease of Installation/Setup**, **Ease of Use**, and **Documentation/Support** criteria are weighted along a continuum, from least to greatest. **Features** were evaluated as present or absent.

Number of Points Assigned to Each Criterion

Ease of Installation/Setup

Knowledge Level Required	
System Administrator	1
UNIX Savvy User	2
Technology Savvy User	3
Knows Basics Only	4
Number of Outside Components Required for Setup	
6+	1
4-5	2
2-3	3
0-1	4

Ease of Use

Prior Understanding of Web Archiving Terminology	
Required	1
Not Required	2
Interface	
Command Line Only	1
GUI Only	2
GUI + Command Line	3

Documentation/Support

Available Documentation for Installation	
None	1
Sparse	2
Adequate	3
Robust	4
Available Documentation for Use	
None	1
Sparse	2
Adequate	3
Robust	4
Online Forum for Troubleshooting	
Not Available	1
Available	2

Features (1 point if present)

- Customization of Crawl
  - Bandwidth Allocation
  - Size
  - Depth
  - Robots.txt Accommodation
  - Dynamic Content Easily Harvested
- Crawl Functionality
  - Crawler Traps
  - Crawl Error Alerts
- Workflow Enhancement
  - Quality Assurance
  - Alterations to Harvested Material
  - Viewer for Captured Sites
  - Metadata Assignment
  - Tracking Permission Status
- Preservation
  - Checksum Assignment
  - Error Detection

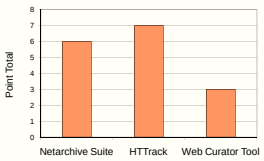
Results

All three tools scored highly and would be good choices for the UARMS setting. Still, each excels in different categories. UARMS may find different tools more appropriate for different stages of a web archiving program or when considering different levels of expertise among those using them.

Overall Scores

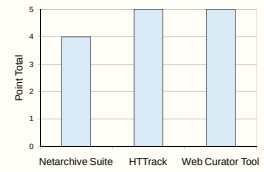
HTTrack	Web Curator Tool	NetarchiveSuite
29	28	25

Ease of Installation/Setup



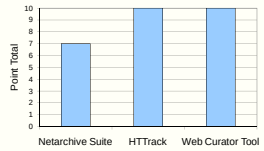
HTTrack is the easiest and quickest tool to install and set up. It can be run on a desktop computer with no specialized components. WCT requires a much higher level of expertise. The Quick Start Version of NetarchiveSuite can be handled by a Unix-savvy user, but the full version (not tested) may require a system administrator.

Ease of Use



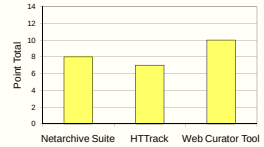
As above, both WCT and HTTrack present a very strong showing when it comes to ease of use. NetarchiveSuite is ranked slightly lower due to its increased use of specialized web archiving terminology.

Documentation/Support



HTTrack has a large amount of online and in-program documentation. WCT offers a number of heavily illustrated and easy-to-read manuals. NetarchiveSuite is more infused with web archiving jargon. All offer a lot of online support, including forums and listservs.

Features



WCT has more features integrated into its program when compared with the other two. This follows from its purpose: to support workflows and quality assurance in addition to harvesting sites. Of the three, only NetarchiveSuite offers preservation functions such as checksum assignment and error checking of files.

Conclusions

At present, HTTrack offers the best solution for UARMS. Because it is strictly a website copier, it does not offer the additional workflow enhancements of WCT or the preservation functions of NetarchiveSuite. However these attributes are less desirable at this point because the UARMS' web archiving program is still young. Instead, staff will benefit from HTTrack's ease of installation, its easy-to-read documentation, and its responsive online forum, while still being able to customize crawls as needed.

WCT may be a good option for the future, because of its user-friendly interface and additional features, like assignment of basic Dublin Core metadata and permissions tracking.