Archiving Michigan State University’s Website: Appraisal, Inventory, and Selection of University Web Properties

By Ed Busch, Julia Corrin, Cynthia Ghering

The Project

Michigan State University Archives & Historical Collections contracted with Internet Archive to use the Archive-It service to crawl, preserve, and provide access to University web sites. Since the university’s web presence is too large to archive in totality, MSU undertook both a self study and a peer review to determine what web content should be archived, identify the frequency of the crawls, and explore existing best practices for web site preservation.

The Internet Archive

Founded in 1996, the non-profit Internet Archive is building a digital library of Internet sites. Their Wayback Machine allows users to see archived version of web pages and their associated data (images, source codes, documents, etc.). With the use of data from Alexa Internet and the Heritrix web crawler, the Internet Archive provides an on-demand archiving service, “Archive-IT,” to institutional subscribers.

The Methodology

The project consisted of eight major activities:
- Literature Review
- Peer review of other web archiving projects
- Online survey of MSU technology staff
- Inventory of the MSU domain
- Creation of a web site collecting plan
- Update of university retention schedule(s)
- Creation of MSU Collections to crawl
- Update of University Archives web site for user access

The first four activities resulted in a comprehensive report, authored by University of Michigan intern Julia Corrin, that offered recommendations for the future of the project and a summation of findings related to the organization, size, and content of the web site. Using the final report as a starting point, University Archives staff developed a web site collecting plan and began archiving MSU web sites and preserving web pages and documents.

What is actually in the MSU Web Domain?

The IT Perspective

A web survey was administered to 800 members of an IT staff email list. From the 67 respondents, it was determined that MSU web properties are substantial and updated frequently. Respondents also indicated that the content of their web sites was, for the most part, not duplicated in print. Staff directories, marketing and current events, policies and procedures and newsletters are posted on many MSU web sites.

Website Inventory

<table>
<thead>
<tr>
<th>There are at least</th>
<th>3,816,112</th>
<th>URLs on the MSU.edu domain.</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are at least</td>
<td>3,670,000</td>
<td>PDFs on the MSU.edu domain.</td>
</tr>
<tr>
<td>There are</td>
<td>1,273</td>
<td>known subdomains.</td>
</tr>
<tr>
<td>That is almost</td>
<td>3,000</td>
<td>URLs per known subdomain.</td>
</tr>
</tbody>
</table>

Of the 439 subdomains, close to 30% were redirected to another subdomain. There were 102 subdomains, or 23%, that included password protected or dynamic content and cannot be archived by current methods.

Retention Schedule Update

- Schedule Number: 116.98
- Schedule Approved Date: General
- Title: MSU Publications
- Disposition: Permanent
- Disposition Description: Retain one copy in office of creation permanently. This office should send copies to archives either as published or on an annual basis. The University Archives is capturing most web sites within the msu.edu domain and identified MSU related external sites. Offices should inform the University Archives before retiring old websites and when creating new web sites.
- Description: These are publications created at MSU including pamphlets, brochures, newsletters, magazines, guide-books, bulletins, programs, announcements, videos, web sites, electronic publications (one-time and serial publications), for on-campus and off-campus audiences. (Revisions in italics.)

Metadata (Dublin Core)

- Title
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Source
- Rights
- Collector

The Results

- Creation of a Web Site Collection Plan (http://archives.msu.edu/documents/CollectionPlan_v1.pdf)
- Update of general retention schedule
- Creation of crawls
  - Administration and Services Collection
  - Colleges, Schools, Research Centers and Institutes Collection
  - Student Organizations and Groups Collection
  - The State News (student newspaper)
  - Topical Events Web Sites Collection
- User access to these collections through the University Archives site, archives.msu.edu, and the Archive-It Partners page

Recommendations

- Collections
  - Refine MSU Collections in Archive-It as we learn more about the MSU web sites.
  - Perform quality checks on captured sites and modify crawls to better represent the web presence.
- Metadata
  - Develop more robust metadata capture. Benchmark other institutions’ metadata models. Add more metadata at Seed and Document levels as practical.
- Outreach
  - Be proactive. Market new web site archiving to MSU colleges and offices. Educate staff during records management training.