MICHIGAN STATE UNIVERSITY

University Archives & Historical Collections

Five Years of Web Archiving @ Michigan State University

Ed Busch

August 18, 2015

Overview

- What We Did
- What We Learned
- What Are We Doing Now
- What We Would Like To Do
- Fun Facts

What We Did

- Our Mission: collect the official records of the institution and preserve the legacy of the nation's pioneer land-grant university.
- Our Goal: To “preserve and make accessible” MSU web sites of enduring historical and research value
  - Almost every office and unit on campus has a web site with business information
  - Content that isn’t preserved anywhere else
    - Publications!
  - Integral to the mission of MSU

What We Did

- Ran trial “snapshots” of msu.edu using Archive-It
- Inventory of MSU related web sites (early 2011)
  - Top level domains = approx. 1,300 sites
  - External domains = approx. 190 sites
    - e.g. coachizzo.com or spartancash.com

What We Did

- Created 3 large collections and 2 smaller special collections
  - Administration and Services; Colleges, Schools, Research Centers & Institutes; and Student Organizations and Groups
  - Topical Events Web Sites; Decommissioned MSU Web Sites
- Added Landing Page to our web site
- Updated Retention Schedule to include web sites
  - Websites are listed with MSU Publications
- Created Web Site Collection Plan
- Added Metadata at collection level
- Identified crawl schedule


MICHIGAN STATE UNIVERSITY
University Archives & Historical Collections
About
Records Management
Collections
Research
Outreach
Giving
Collections
Guides
On the Banks of the Red Cedar
Exhibits
Documenting Diversity @ MSU
Civil War
Vietnam Project Archives
Web Archives
Home > Collections >Web Archives
Web Archives @Michigan State University
Early attempts at preserving copies of the MSU Web Sites began by UAHC staff in 1998 and continued through 2006. In parallel, the Internet Archive began crawling some of the MSU web sites in 1997 and continued until 2009.
Beginning in January 2011, MSU initiated a partnership with the Internet Archive's Archive-It program to collect and preserve the University's web pages of historical value produced by administrative offices, schools, departments, service units, institutes, centers, programs, and faculty, student and alumni organizations. Access to these archives is provided by the Internet Archive's Wayback Machine technology. Our Web Site Collection Plan (PDF)* documents the methodology for selecting and preserving MSU sites.
Access to the collections of MSU websites is:
1998-2006 (UAHC crawls) Available only at the UAHC
1997 – 2009
2011 -> (These may be searched directly using the search box to the right or by the Archive-It interface.)
Search MSU Web Archives
Go
Frequently Asked Questions

What We Learned

- Once you create a collection, you can't split or combine easily
  - What's the best collection creation strategy- to lump or not? Smaller seems to provide more flexibility.
  - I've kept new collections smaller

What We Learned

- Our New Collections
  - Michigan State University Libraries Collection
  - MSU Administration and Services Collection*
  - MSU Alumni and Fan Sites Collection
  - MSU Athletics Collection
  - MSU Colleges, Schools, Research Centers & Institutes Collection*
  - MSU Employee Unions Collection
  - MSU Related News Publications Collection
  - MSU Social Media Collection
  - MSU Sponsored Projects Collection
  - MSU Student Organizations and Groups Collection*
  - MSU Topical Events and Subjects Web Sites Collection
  - MSU Arts and Culture Collection

What We Learned

- Some sites are just difficult to crawl – "recursive issues"
  - Using regular expressions and constraints – Archive-It staff very helpful
  - Text matching
  - Lots of test runs – takes time

What We Learned

- Web Archiving requires more staff time than expected
- Websites are being created or modified every day
  - New functionality often causes problems in next crawl
  - Lots of sites from outside the domain are related to the university.
  - Units reorganize and change domains
  - Run Test crawls!

What Are We Doing Now

- Have captured over 4.2 TB since 2011.
  - 668 active seeds
- Revisiting problem seeds
- Spending more time on quality check
  - Particularly prime websites like president's office, board of trustees, etc.
  - Getting feedback from users on a few captured sites

What Are We Doing Now

- Member of campus Web Development group (WebDev CAFÉ) – meets monthly
  - Gives me an insight on what's going on with university websites and to remind them to let me know when a site is changing or being removed.
- New sites
  - Social Media sites
  - Non-MSU Historical Collection sites
- Working with users who need permanent links to content
  - One professor has been working with me to get faculty retiree biographical information captured

What We Would Like To Do

- Expand Social media crawls
- Catalog our Crawl Collections
- Investigate using data for Digital Humanities research
- Add more metadata
  - Need to decide what would be useful

Fun Statistics

- Google Search Operator statistics
  - site:msu.edu – 5,910,000, site:umich.edu 12,600,000
  - Shows number of indexed URLS – includes duplicates
  - site:msu.edu filetype:pdf – 1,040,000
  - site:umich.edu filetype:pdf – 669,000
- www.check-domains.com
  - msu.edu Google indexed pages – 427,000
  - umich.edu Google indexed pages – 1,830,000

Wrap Up

- Questions, Comments, Suggestions?
- Contact
  - Ed Busch
  - Electronic Records Archivist
  - buschedw@msu.edu


MICHIGAN STATE
UNIVERSITY
University Archives
&
Historical Collections