Web Site Collection Plan

for

Michigan State University Archives
& Historical Collections

May 21, 2015

Prepared by:

Ed Busch
Michigan State University
buschedw@msu.edu

Contents

Section 1. Overview, Mission, Vision & Scope

A. Overview

The mandate of University Archives & Historical Collections (UAHC) is founded on a resolution of the Board of Trustees, as recorded in the minutes of November 21, 1969. This resolution claims all records of the official activities of university officers and offices are the property of Michigan State University, and that such property cannot be destroyed without the approval of the director of Archives. University units increasingly utilize Websites to document their actions as well as to publicize their policies and activities. Electronic records management applies the same principles of paper records management in an electronic environment. Many of the traditional paper records now are found only in an electronic or digital format.

Early attempts at preserving copies of the MSU Websites 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 2010. Beginning in January 2011, MSU initiated a subscription with the Archive-It service to collect and preserve the Websites of historical value and the Internet Archive's Wayback Machine technology for access to these pages.

B. Mission Statement

University Archives & Historical Collections (UAHC) provides records management services to the university and preserves and provides access to the institution's historical records. The UAHC also maintains historical collections that support faculty and student research and classroom instruction.

C. Vision Statement

The vision of University Archives & Historical Collections is to:

- Provide departments and campus units with efficient and timely records management services
- Expand the services of the records management program to include digital and born digital records and documents
- Deliver high quality responses to administrative, departmental, and research requests for information
- Encourage research collections to be incorporated into the teaching experience to provide unique educational and research opportunities for MSU students
- Provide access to the UAHC's collections through the reading room, online resources, and social networking

D. User Groups

As the official repository of Michigan State University's permanent records, UAHC serves the entire university community including its administration, faculty, staff, and students. It supports and encourages new research by scholars from MSU and from other institutions as well as the general public.

The UAHC welcomes comments, questions, and suggestions in regards to MSU Web Archives. Contact the UAHC at archives@msu.edu.

E. Collection Subject, Theme, or Event

UAHC is the official and foremost repository for records pertaining to the history of Michigan State University. As such, the UAHC seeks to collect and preserve the MSU Websites of historical value and provide access to these pages.

F. Web Archives Curator(s)

Ed Busch
Electronic Records Archivist
University Archives & Historical Collections
Michigan State University
East Lansing, MI 48824
buschedw@msu.edu

Section 2. Selection

A. Sources

Because of its stated commitment to capture, preserve and provide access to University Websites, the staff of the UAHC appraises Websites to determine if they qualify as a digital asset appropriate for crawling. Potential Websites are appraised to determine if UAHC should preserve the content. This is based on practical limitations and the stipulations of our UAHC resources to collect, store and manage the content.

Web sites selected for inclusion are:

- Sites (URLs) containing official MSU information, either hosted on university Web servers or hosted by private companies. In general, captured Websites will contain only university information and do not contain a combination of public and private information. However, some private Websites may be considered for capture if they contain significant university information and/or assist in the formation of university policy.
- Sites (URLs) related to UAHC held historical collections and those of geographically local significance.

The UAHC has identified the following MSU collections within the MSU Web Archives:

- MSU Administration and Services Collection
- MSU Colleges, Schools, Research Centers, and Institutes Collection
- MSU Student Organizations and Groups Collection
- MSU Topical Events Web Sites Collection
- Brain Images Web Sites Collection
- MSU Libraries Websites Collection
- MSU Alumni and Fan Sites Collection
- MSU Arts and Culture Collection
- MSU Athletics Collection
- MSU Employees Unions Collection
- MSU Related News Publications Collection
- MSU Social Media Collection
- MSU Sponsored Projects Collection

There are two historical Websites collections:

- Lansing Area Historical Websites Collection
- UAHC Historical Collections

There are also two ad hoc collections that are currently inactive:

- Decommissioned MSU Web Sites
- MSU PDFs

The seeds included in these collections can be found on the Archive-It website for the UAHC collections at http://archive-it.org/home/MSUArchives/?show=Collections.

B. Capture Scope

The Web sites crawls are based on a list of URLs, or a seed list developed by the UAHC curator. The curator identifies the name of each site and briefly describes the sites using metadata. This provides an overall descriptive identifier. In general the curator makes an effort to maintain the organic unity of the site in the archive, as this is important in order to

capture the look and feel. Embedded images and style sheets are crawled. In some cases, certain web pages that include confidential information or information that the content provider does not want to be archived will not be crawled.

Social media websites frequently change. The goal of the UAHC crawls of social media websites is to capture the MSU content and not necessarily the look and feel of the social media site. Primary focus is on the ability to view the MSU content.

C. Rights Management

1. Copyright Compliance

The UAHC will comply with all copyright laws applicable to captured and archived materials. Copyright statements attached to digital materials must remain intact as these materials are harvested, ingested, and deposited in the digital archive.

1.1 Copyright Ownership and U.S. Materials

Copyright U.S. materials are selected from the free Web and deposited in the archive under a claim of fair use; however, if UAHC receives notice from a copyright holder challenging the legality of holding a protected item in the digital archive, the material will be promptly removed from the archive and re-harvested only upon reaching a resolution with the copyright holder.

1.2. Copyright Ownership and Foreign Materials

To ensure compliance with international copyright laws, written permission will be sought before archiving and providing access to digital materials created in foreign countries and protected by foreign rights laws.

2. Rights Statements

Educational use only, no other permissions given. The UAHC does not assert ownership rights over the intellectual property of the contents included in the web archive collection. All rights of ownership remain with the owner(s) identified on the Website for the full term of copyright. The UAHC is not involved in the creation of the Websites and has no oversight for the contents of these collected Websites. The UAHC assumes no responsibility for the accuracy or lawfulness of the collected Websites or the contents within. These captured websites are provided here for educational purposes. They may not be reproduced or distributed in any format without written permission of the owner(s).

Section 3. Acquisition

A. Frequency of Capture

The frequency of Website changes and revisions may vary from campus unit to unit. Some units that publish fairly static versions of policies, publications, or images may not change their Website design for months at a time. However, units with more advanced services (interactive forms, streaming video, etc.) may undergo major changes numerous times each year. Recognizing these differences, UAHC will crawl a Website at least once a year, and more frequently for those sites that change often. Our default crawl frequency will be quarterly.

Note: It is important to note that sometimes a Website will change while it is being harvested. Pages may be removed, moved, or may be changed by the content provider during the capture. This could result in hyperlink errors or semantic inconsistencies among pages of a captured web site when subsequently viewed by users.

B. Capture Scope

Capture scope and URLs on the seed list will be re-evaluated and adjusted on a periodic basis based on the results of each capture and changes to the source Websites.

C. Material Types & Formats

Since it is important to capture the “look and feel” of the Websites, the UAHC will capture images, audio, and video files along with text. Captured format types include: text, jpeg, xml, pdf, gif, javascript, mpeg, png, flash, msword, quicktime, mp4, and many others. An effort will be made to capture interactive pages. Embedded databases may not be able to be captured for technical reasons.

D. Interactive & Dynamic Content

Websites will be evaluated for interactive and dynamic content. If a site is password protected, it will not be captured without approval. Email links and comment forms may not be included.

It is very important to retain the interactive and dynamic functionality of the original Website since this project not only seeks to capture content but the artifact, which is why the capture setting is “host + linked pages.”

Section 4. Descriptive Metadata

A. Level of description

Basic metadata will be provided at the collection level: title, creator, subject, description, publisher, date, type, coverage, rights and collector.

At the seed level, an effort will be made to provide title and creator metadata. Currently, no effort will be made to provide metadata at the document level.

B. Metadata elements

There are no plans at this time to provide customized metadata.

C. Controlled vocabularies

The UAHC uses controlled subject headings as well as locally defined authority lists.

Section 5. Presentation and Access

A. Discovery

Collections are automatically indexed as part of the Archive-It subscription. Users will be able to search for archived items of interest using keywords through the archive-it.org partners search or URL search through the Internet Archive waybackmachine.org site or a search function at the UAHC Web Archives *http://archives.msu.edu/collections/webarchive.php?collections_webarchive*.

In addition, Archive-It provides a public interface to browse through an alphabetical list of all URL's. At the bottom of every public web page, users can also find out which Archive-It partner is crawling the URL, the collection the URL is assigned to, and the frequency at which the URL is being crawled. By default, seed title and description metadata will be displayed in the Public URL Registry.

B. Access

Crawled collections will be visible to all users with a connection to the internet. There is no plan to collect restricted or private information during our Website crawls.

C. Look-and-Feel

Preserving “look and feel” is an essential element of this project. Every effort will be made to avoid disassociating information from its context. This effort may be hampered by web authors use of the robots.txt file and advanced website programing.

D. Dynamic Content

It is important that users be a given access to hyperlinks to materials outside the Web archive, but UAHC will not ensure that such links are active. The UAHC will attempt to maintain all Website functionality but may be limited to available crawling technology.

E. Multiple Types/Formats

Whenever possible information objects will be preserved and made accessible in formats as they are on the original Website.

F. Authenticity

The authenticity of items archived by the UAHC is established by the Archive-It interface, which features a banner marking the items with “You are viewing an archived web page, collected at the request of Michigan State University using Archive-It. This page was captured on HH:MM:SS Mmm DD, YYYY, and is part of the “collection_name” collection. The information on this web page may be out of date. See All versions of this archived page.”

Section 6. Maintenance and Weeding

A. Maintenance Activities

Seed lists and capture parameters will be updated periodically. Descriptive metadata and rights management will only be updated when new information is obtained.

B. Deselection Guidelines

In most cases deselecting or weeding will be undertaken according to the guidelines that govern archival collection maintenance. This is an ongoing project and from time to time the curator will need to evaluate the Web archive in light of changing scholarship and user demand. The curators may determine that certain Websites should be “de-accessioned.”

Currently, the Archive-It subscription service does not contain subscriber functionality to remove or merge collections.

C. Collection Evaluation

Usage, researcher feedback, and an ongoing evaluation of the scholarly literature may lead to adjustments in collection guidelines. Collection evaluation may include usage statistics, user feedback, citation analysis, and/or surveys of and assessments by subject specialists. The evaluation may result in the curator moving seeds between collections, creating new collections or making a collection inactive or dormant.

Section 7. Preservation

A. Technology Obsolescence

For the storage, preservation, and management of its digital collections, UAHC utilizes the Archive-It subscription. They store two copies online with regular integrity checks and work with partners to have redundant copies in other locations at the Bibliotheca Alexandrina in Egypt and other locations in the U.S. In addition, a copy of the data will be sent yearly to UAHC for local use and preservation.

Captured files are stored in the WARC archival file format.

B. Preservation Metadata

The Archive-It subscription software provides (for each web crawl collection) 15 Standard Dublin Core Metadata fields and the capability to create custom metadata fields. These fields are:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Relation
- Coverage
- Rights
- Collector

The software currently only requires the Description field.

Additionally, the software provides the same metadata fields for seeds (URLs) and documents (captured web pages).

Appendix A. Web Archiving Service Agreement

TERMS AND CONDITIONS FOR ARCHIVE-IT
Use of Service. Subject to the terms and conditions of this Agreement, Internet Archive hereby grants to Institution during the Term a non-exclusive, non-transferable right to access and use the Service to create and use Collections, as further described and defined in Exhibit A, solely for Institution's own internal purposes, Institution acknowledges and agrees that Internet Archive retains all right, title and interest in and to the Service and all intellectual property rights contained therein. All rights not expressly granted to Institution under this Agreement are expressly reserved to Internet Archive.

Access of Service by Users. Institution may authorize up to twenty (20) individuals to access the Service ("Users"). Users must be employed faculty, researchers, and/or staff at the Institution or officially affiliated researchers. The Institution may transfer access privilege among Institution employees or affiliated researchers so long as the total remains below 20 individuals. Internet Archive will provide Institution with a unique username and password to enable Institution to access the administrator account for the Service. Institution is responsible for establishing usernames and passwords for its Users and for maintaining the confidentiality of all usernames and passwords, and is solely responsible for all activities that occur under any username, including, but nor limited to, the username for the administrator account. Institution agrees (a) not to allow a third party to use its administrator username or password at any time; and (b) notify Internet Archive promptly of any actual or suspected unauthorized use of any usernames or passwords, or any other breach or suspected breach of this Agreement. Internet Archive reserves the right to terminate any username and password, which Internet Archive reasonably determines may have been used by an unauthorized third party or by any User or individual other than the User to whom such username and password was originally assigned.

Creation of Collections. Institution will be solely responsible for creating, reviewing, editing, and otherwise controlling any Collections. Institution acknowledges that Internet Archive has complete discretion over the Collections made available by Internet Archive on the Website that is described and defined in Exhibit A, and that Internet Archive has no obligation to Institution or any third party, and undertakes no responsibility, to review the Collections to determine whether the Collections may incur liability to third parties. Notwithstanding anything to the contrary in this Agreement, if Internet Archive believes in its sole discretion that any materials in the Collections may create liability for Internet Archive or Institution, Internet Archive has the right to take any actions with respect to the Website and the Collections to minimize or eliminate this liability, including, but not limited to, removing any or all Collections from the Website. Internet Archive will inform Institution of such actions, but will not be required to get any approval from Institution or any third party, neither before, nor after taking such actions. Some of the data collected and made available in Collections through the Service may be governed by local, national, and/or international laws and regulations, and the use of such content by Institution is solely at Institution's own risk.

Availability of Service and Website. Internet Archive will use reasonable commercial efforts to make the Service available to Institution and to host the Website on Internet Archive's or its contractor's servers and to make any Collections created under this Agreement publicly available to anyone on the Website. Internet Archive will use reasonable commercial efforts to inform Institution and User if the Website and/or Service is not available. Internet Archive will provide electronic Help Desk support to Users via a dedicated e-mail address.

Fees and Payment. Institution will pay to Internet Archive the annual subscription fees set forth in Exhibit A. Except as otherwise provided in Exhibit A, all fees and other charges are due and payable to Internet Archive within sixty (60) days after the date of Internet Archive's invoice. If exempt from sales tax, Institution shall provide Internet Archive with a copy of its exemption certificate. If Institution is not able to prove exemption to Internet Archive, Institution shall be responsible for payment of any sales, use, value-added and other taxes, and all fees or assessments levied by any governmental body to which Institution's payments with respect to the Services may be subject.

Term and Termination. This Agreement commences on the Effective Date and remains in effect for a period of twelve (12) months (the "Initial Term") unless earlier terminated as set forth below. Internet Archive may adjust the fees to be paid by Institution for a renewal upon issuing an updated Exhibit A to Institution at least sixty (60) days before the beginning of such renewal. The Initial Term and any renewals thereof shall be Collectively referred to as the "Term", Either party may terminate this Agreement (a) for convenience upon thirty (30) days written notice to the other party; or (b) immediately upon written notice to the other party if the other party materially breaches this Agreement, and such breach remains uncured more than thirty (30) days after receipt of written notice of such breach. Internet Archive may suspend or terminate Institution's access to the Service at any time without prior notice in order to: (a) prevent damages to, or degradation of, Internet Archive's Internet network integrity; (b) comply with any law, regulation, court order, or other governmental liability resulting from Institution 's use of or access to the Service. Upon suspension, termination or expiration of this Agreement for any reason all rights and obligations of both parties, including all rights granted hereunder, shal1 immediately terminate. Sections 3, 0, 6, 7 and 8 and any payment obligations incurred prior to the suspension, termination or expiration shall survive such suspension, termination or expiration of this Agreement. If the Agreement is terminated for convenience by the Institution as set forth in Section 0 during a Term, Internet Archive will refund to Institution a pro-rata amount of the annual subscription fee paid for the unused portion of such Term. If the Agreement is terminated for convenience by the Internet Archive, the Internet Archive shall refund two times the amount of the annual fee pro rata fee. After suspension, termination or expiration of this Agreement, Internet Archive may move any or all Collections created under this Agreement to a general archive operated by Internet Archive and may make it publicly available by hosting it on a website without any further obligations, including, but not limited to, any obligation to pay any royalties or fees, to Institution. The Institution, at time of suspension, termination or expiration, may request that public access be removed.

1. Disclaimer of Warranties. THE SERVICES DESCRIBED IN THIS AGREEMENT ARE PROVIDED TO INSTITUTION ON AN "AS IS" BASIS WITHOUT WARRANTIES OF ANY KIND. WITHOUT LIMITING THE FOREGOING, INTERNET ARCHIVE DISCLAIMS ALL WARRANTIES AND REPRESENTATIONS OF ANY KIND, WHETHER EXPRESS, IMPLIED, OR STATUTORY, INCLUDING WITHOUT LIMITATION THE IMPLIED WARRANTIES OF MERCHANTABILITY, TITLE, NON-INFRTNGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE. INTERNET ARCHIVE DOES NOT WARRANT THAT THE OPERATION OF THE SERVICE AND/OR THE WEBSITE SHALL BE UNINTERRUPTED. IN THE EVENT OF AN INTERRUPTION OF THE OPERATION OF THE SERVICE AND/OR THE WEBSITE, INTERNET ARCHIVE'S SOLE OBLIGATION SHALL BE TO RESTORE OPERATION AS SOON AS REASONABLE POSSIBLE.

2. Indemnification. Institution shall, upon request, indemnify Internet Archive and its affiliates, and each of their respective employees, directors and representatives, from and against any and all claims, costs, losses, damages, liabilities, judgments, penalties and expenses (including reasonable fees of attorneys and other professionals), arising out of or

related to (a) Institution's and its Users' use of and access to the Service, including any Collections generated from the use of the Service; or
(b) any improper or unauthorized use of the Service by Institution or its Users.

8. General. Institution may not assign this Agreement to any third party without Internet Archive's prior written consent. Any purported assignment in derogation of the foregoing will be null and void. All notices provided under this Agreement will be in writing and will be delivered by personal delivery, private courier, or by certified or registered mail, return receipt requested, to the addresses set forth in this Agreement, and shall be deemed given upon receipt (or when delivery is refused). No waiver of any terms or conditions of this Agreement will be valid or binding on a party unless such party makes the waiver in writing. This Agreement may not be altered, amended, modified, or otherwise changed in any way except by a written instrument signed by the authorized representatives of each party. If any provision of this Agreement is found or held to be invalid or unenforceable, then the meaning of such provision will be construed so as to render the provision enforceable, and if no feasible interpretation would save such provision, it will be severed from the remainder of this Agreement, which will remain in full force and effect. Neither party shall be considered in default of its performance under this Agreement to the extent such performance is prevented, restricted or interfered with by any act or condition whatsoever beyond the reasonable control of such party. Institution's relationship to Internet Archive is that of an independent contractor, and neither party is an agent or partner of the other. This Agreement may be executed in counterparts, each of which will be considered an original, but all of which together will constitute one and the same instrument.

Exhibit A

1. Description of Service

1.1 The "Archive-It Service". Internet Archive has developed a web application to provide web archiving services for library, archive and other non-profit organizations. This service, named "Archive-ft", is a web-based application designed for institutions interested in archiving, accessing and managing web content without needing specific technical expertise or resources (the "Service").

1.2 Collections. The Service allows Users to login and create web collections ("Collections"), catalogue the website;:s associated with a Collection, archive websites in the Collection, monitor the archiving process, search and browse the Collections when complete, and access these Collections. Internet Archive will host and manage any and all Collections created under this Agreement for the duration of this Agreement and will make them publicly accessible.

1.3 Website Access to Collections. Internet Archive will make the Collections publicly available under the following URL: httpllwww.archive-it.org (the "Website"). Each Collection will be accessible via a text search engine and a tool to render archived webpages (i.e., the Wayback machine interface). On a case by case basis, access to Collections can be restricted to a defined group of people for a definitive period as agreed upon between Internet Archive and Institution.

1.4 Management of collected Data. All data and information collected through the Service will be stored at one of Internet Archive's data centers. The Collections created through the Service and associated content from websites will be kept in a separate repository from Internet Archive's general web archive during any Term. A primary and backup copy of this data will be made available online. A copy of the data collected for the Institution will be made available up to two (2) times per year during the subscription period via hard drive, or Internet transfer for an additional fee.

2. Fees

2.1 Subscription

- Crawls: Maximum of Three Hundred (300) seeds across one to three (I -3) active Collections
- Ability to customize frequency of Crawls
- Up to Twelve Million (12,000,000) URLs archived
- Not to exceed One Point Zero (1 .0) terabytes in data

"Crawls" means the content associated with a web capture operation that is conducted by a User.
"Seeds" means the initial starting point of a Crawl in the process of creating a Collection.