WEB ARCHIVE COLLECTING POLICY

Purdue University Libraries
Virginia Kelly Karnes Archives and Special Collections Research Center
504 West State Street
West Lafayette, Indiana 47907-2058
(765) 494-2839

http://www.lib.purdue.edu/spcol

Created by: Ankur Bhatia, December 17, 2013.
Revised by: Carly Dearborn, November 19, 2015

Contents

Section 1. Overview, Mission, Vision & Scope 3

Section 2. Selection.....4

Section 3. Acquisition 4

Section 4. Descriptive Metadata 5

Section 5. Presentation and Access.....5

Section 6. Maintenance…….....6

Section 7. Reference 6

Section 1. Overview, Mission & Scope

A. Overview

Increasingly, many vital university records are both created and disseminated online. These records not only detail major Purdue activities such as the creation of new departments, construction of new buildings, major campus initiatives, or changes to the curriculum but also provide valuable historical context for what the university was like at a given place in time.

The very nature of the web places many of this content at risk. Web sites are abandoned, links become corrupt, and campus departments restructure. Capturing a web site at a specific point in time, when it is at the peak of its use, is a useful way to save the web site as a medium but also to save the informational content. The Virginia Kelly Karnes Archives and Special Collections recognized this need to preserve Purdue's ephemeral records. In June 2012, the Archives, with support from the Office of the Provost and in participation with Libraries Digital Programs, established the Purdue University Web Archive. Initially established utilizing the California Digital Library's Web Archiving Service (WAS), the web archive transferred platforms to the Internet Archive's Archive-It Service following the WAS-Archive-It partnership in 2015. The Archives actively collects, preserves, and ensures access to Purdue's digital publications, records, and web culture.

B. Mission Statement

The mission of the Purdue University Libraries Archives and Special Collections is to support the discovery, learning, and engagement goals of Purdue University by identifying, collecting, preserving, and making available for research records and papers of enduring value created or received by the University and its employees. As part of the Archives and Special Collections, the Purdue University Web Archive also follows this mission.

C. User Groups

The primary audience for The Purdue University Web Archive will be students, faculty, staff, alumni and researchers interested in Purdue University history or land grant institutions.

D. Collection Subject, Theme, or Event

Archives and Special Collections is the official and foremost repository for records pertaining to the history of Purdue University. As such, the Purdue University Web Archive collects Purdue web pages of historic value and provides access to these pages.

Section 2. Selection

A. Criteria

Archivists select content for the Purdue University Web Archive based on the following criteria:

- Materials relate to the history, administration, or culture of Purdue University (University Archives)
- Materials relate to a subject area of distinction for Purdue University as the land grant university for the state of Indiana (agriculture, engineering, science, technology)
- Materials are rare or unique and support the research and teaching needs of the University
- Materials complement the existing collection's strengths or areas of subject emphasis
- Material, produced by the University traditionally in print, but that are now only published digitally

B. Resource Constraints

There are several resource constraints: 1) cost associated with storage space, 2) maximum storage space available. Judicious use of harvesting scope will result in more captures in same amount of space.

Section 3. Acquisition

A. Capture Scope

The initial crawl of web sites will be based on a list of URLs, or a seed list developed by the Digital Preservation and Electronic Records Archivist. The archivist will identify the name of each site and briefly describe the site. This will provide an overall descriptive identifier. In general the archivist will make 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 will be 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 (Robot Exclusion). All jobs will default to "max-link-hops = 25" scope setting. This means that the crawler will follow links from the URL(s) entered for the site, and continue gathering files until it gets 25 links away from the starting point. This should provide a thorough capture of most sites.

B. Frequency of Capture

The frequency of web site capture may vary from case to case. Web sites that contain comparatively static publications, images, etc. may not change their content very often and hence these web sites will be crawled once a year until any noticeable changes are observed in the website. However, web sites that undergo major changes frequently will be crawled more

often. The frequency of crawls for such web sites will be in sync with the frequency of changes made to the site.

C. Material Types & Formats

Efforts will be made to ensure the ‘’look and feel’’ of the original web site is maintained to the greatest possible extent. In this regard various file formats such as HTML, PDFs, Office Document, Images, Audio, Video or Compressed will be part of the harvest. Static, dynamic and interactive pages will be captured as part of the harvest.

Section 4. Descriptive Metadata

The Archives will create descriptive metadata using Dublin Core’s 15 standard fields. Dublin Core is used widely for describing and cataloging digital materials and is managed by the Dublin Core Metadata Initiative®. Collection and seed level metadata is available to harvesters, including OCLC’s WorldCat catalog.

Section 5. Presentation and Access

A. Look and Feel

All possible and technically feasible efforts will be made to ensure that the crawled web site resembles the original web site closely. The only exception to this will be the items on the web site that are protected by the web authors using robots.txt file.

B. Access

Archived sites are available to view within 24 hours after a crawl has been completed; however, full-text searching can take up to 7 days to finish processing. Users can access Purdue’s web archive by browsing through Archive-It.org or through Purdue’s landing page at https://www.archive-it.org/organizations/979. All archived sites will appear with a header so as not to be confused with the original web site.

C. Authenticity

The authenticity of the crawl is established with the banner on the top of the crawl featuring the original website URL, the archived URL and the date captured. All archived web sites are tested following the complete of the crawl to ensure the site has been captured properly.

Section 6. Maintenance

A. New Web Content

The Archives is committed to investigating new content to contribute to the web archive.

B. Ongoing Maintenance Activities

The harvests of existing web sites will be revisited every quarter to see if there is significant change in the website to initiate re-harvest. Descriptive metadata of existing and new crawls will also be kept updated with new information.

C. Quarterly Evaluation

The current list of harvested websites will periodically be revisited and potential additions to this list will be discussed. Feedback from the researchers and scholars if deemed fit will be accounted for in the next quarter.

Section 7. Reference

The following policies were referred during preparation of Purdue University Web Archive Collecting Policy.

Michigan State University Collection Plan, Michigan State University Archives & Historical Collections, http://archives.msu.edu/documents/CollectionPlan_v1.pdf

Northwestern policy on web archiving, Northwestern University Library, http://www.library.northwestern.edu/webarchives

Tamiment Web Archiving Collecting Policy, The Tamiment Library & Robert F. Wagner Labor Archives, New York University Libraries, http://www.nyu.edu/library/bobst/research/tam/tam_web_collecting_policy_2010-09-08.pdf