IMLS Award Number LG-06-09-0174-09

Classification of the End-of-Term Archive:

Extending Collection Development Practices to Web Archives

INTERIM PERFORMANCE REPORT

July 2010

Submitted by:

Cathy N Hartman

Cathy Nelson Hartman
Principal Investigator
940-565-4369
cathy.hartman@unt.edu

Kathleen Murray
Senior Research Fellow and Project Coordinator
kathleen.murray@unt.edu

University of North Texas
UNT Libraries
1155 Union Circle #305190
Denton, TX 76203-5017

Introduction

This is the first interim performance report for the project titled: *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives.* The reporting period is December 1, 2009 – June 30, 2010.

The project is comprised of the two work areas: Archive Classification and Web Archive Metrics (Figure 1). This report includes three sections: Goals and Accomplishments, Significant Findings and Accomplishments, and Project Achievements and Lessons Learned.

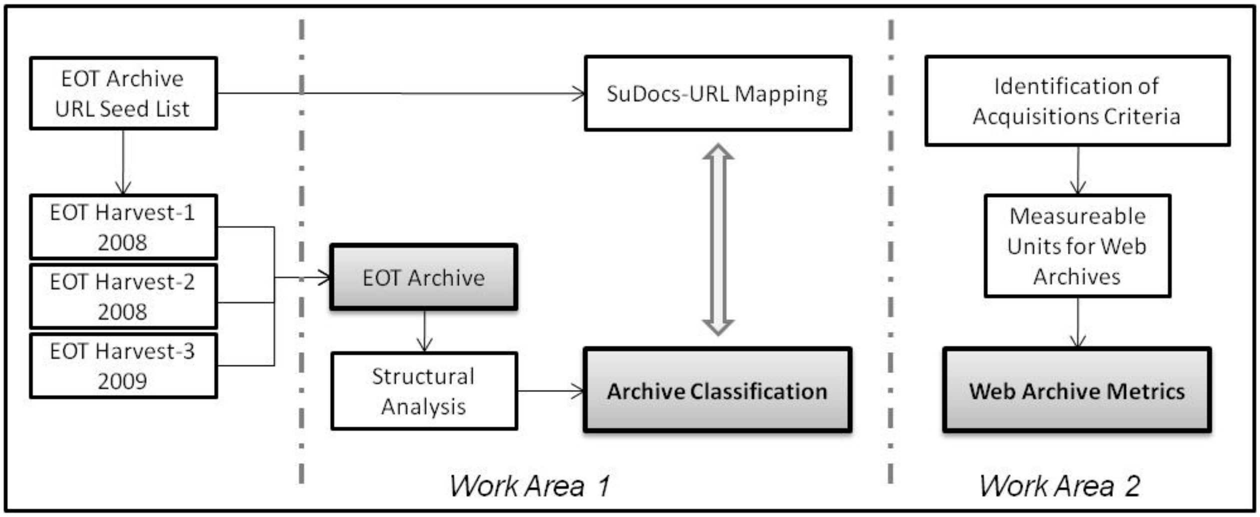

Figure 1. Project Work Areas

I. Interim Goals and Accomplishments

A. Archive Classification

1. Data Storage Environment for Research and Testing
    - Purchase and installation of equipment complete
    - EOT Archive transferred, installed, and validated
    - Development environment created

2. Structural Analysis
    - Link extraction and analysis of archive content completed
    - On-site consultation with Internet Archive staff regarding current archive analysis tools

B. Web Archive Metrics

1. Identification of Acquisitions Criteria
    - Preliminary analysis of library statistics and measurements: ACRL, ARL, & NCES
    - Focus group discussion conducted and transcribed

- License for Web survey creation tool acquired. (*Note*: A survey of depository libraries is planned in the fall of 2010. This has been added to the project to assess libraries' needs for acquisition versus access in regard to the EOT Archive.)

II. Significant Findings & Accomplishments

1. Archive Classification
   a. EOT Archive (c. 16 TB) transferred from Library of Congress to static servers at UNT
   b. File format verification completed
   c. *Wayback Machine* interface to the Archive created
   d. EOT Archive Statistics: Total URIs = 160,156,233

Domain	# URIs	# Unique Sub-domains
gov	137,780,023	14,338
com	7,805,205	57,873
org	5,107,552	29,798
mil	3,554,956	1,677
edu	3,551,845	13,856

Table 1. Number of URIs & Subdomains by Domain

Mimetype	# Files
text/html	105,590,929
image/jpeg	13,665,196
image/gif	13,031,046
application/pdf	10,320,163

Table 2. Number of Files by Mimetype

   e. Given the scale of the Archive, a decision was made to include two domains in the classification:
      i. .gov
      ii. .mil

2. Web Archive Metrics
   a. "Cost per use" is a critical metric for libraries. It drives acquisitions/retentions. As libraries increasingly move their collections to digital formats, meaningful measures of use become more fundamental to collection management decisions.
   b. Measure of a library's "owned" collection, by traditional measures of "how much" and "how many", become less representative of the scope of overall access to materials a library provides via licensing and consortia arrangements.
   c. Discovery of new subject-relevant, web-published resources is a critical selection issue for librarians charged with collection development.
   d. The top three types of digital content participants select for the collection(s) they manage are:

    i. web-published reports
    ii. agency/organizational websites
    iii. statistical databases
  e. GPO 's 2009 Biennial Survey of Federal Depository Libraries & Library Needs Assessment
    i. Question 18b: Are you interested in receiving digital files [of online publications] on deposit? [Yes=30%]
    ii. Project SMEs suggest that number may be high and indicate that libraries are moving toward depending on a trusted source to provide permanent public access rather than building their own collections. Because of this disparity, this project plans a survey of Federal Depository Libraries in the fall 2010 to clarify responses to this question on the GPO 2009 Biennial Survey
    iii. Critical implications for collection building vis-à-vis the EOT Archive:
      1. Services needed
      2. Cost model

III. Project Achievements & Lessons Learned

Achievements related to project management and communications are listed below.

1. Institutional Review Board (IRB)
   a. Protocols developed for focus group discussions
      i. Discussion guides
      ii. Questionnaires
   b. Approval from UNT IRB granted
2. Advisory Board
   a. Established board comprised of national and international leaders in the area of Web archives
   b. Two meetings with the board were held:
      i. December 2009 in Washington DC
      ii. June 2010 conference call
3. Subject Matter Experts
   a. Recruited 10 SMEs in the area of government information
   b. First meeting: April 25, 2010 in Buffalo, NY
4. Project Staff
   a. Graduate Research Assistants hired: Assessment & Programming
   b. Regular team meetings held
5. Project Web Presence
   a. Project wiki created
   b. http://research.library.unt.edu/eotcd/
6. Presentations
   a. Murray, K. R., Phillips, M., & Hartman, C. N. (2010, April 25). *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives.* Presented at the SME Meeting in Buffalo, NY. Available: http://research.library.unt.edu/eotcd/w/images/1/12/SME_Meeting_20100425_Buffalo_NY.pdf

   b. Phillips, M. (2010, May 3). *Classification of the End-of-Term Archive.* Presented at the International Internet Preservation Consortium General Assembly 2010 in Singapore.

7. Conference Attendance
   a. Web Wise 2010 Conference, Denver, CO, March 3 – 5, 2010.
      http://www.bcr.org/webwise2010/
   b. International Internet Preservation Consortium (IIPC) General Assembly, National Library, Singapore, May 3 - 7, 2010
      http://netpreserve.org/events/singapore.php
      i. Metrics Working Group: Report expected in Fall 2010 to include three classes of metrics for Web archives: Counts (size, numbers); Access (discovery, usage); Costs
   c. IS&T Archiving 2010 Conference, National Library, Den Haag, the Netherlands, June 1-4, 2010
      http://www.imaging.org/ist/conferences/archiving/Archiving%202010%20Preliminary%20Program.pdf