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

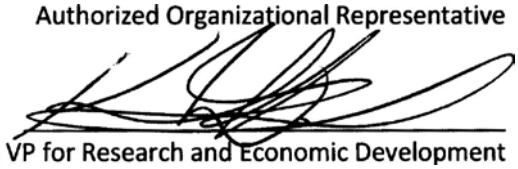
December 23, 2010

Submitted by:

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

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

Authorized Organizational Representative



Kenneth Sewell, Associate VP for Research and Economic Development

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

Introduction

This is the second interim performance report for the *eotcd* project, which is formally titled *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives.* The current reporting period is July 1, 2010 – December 31, 2010.

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

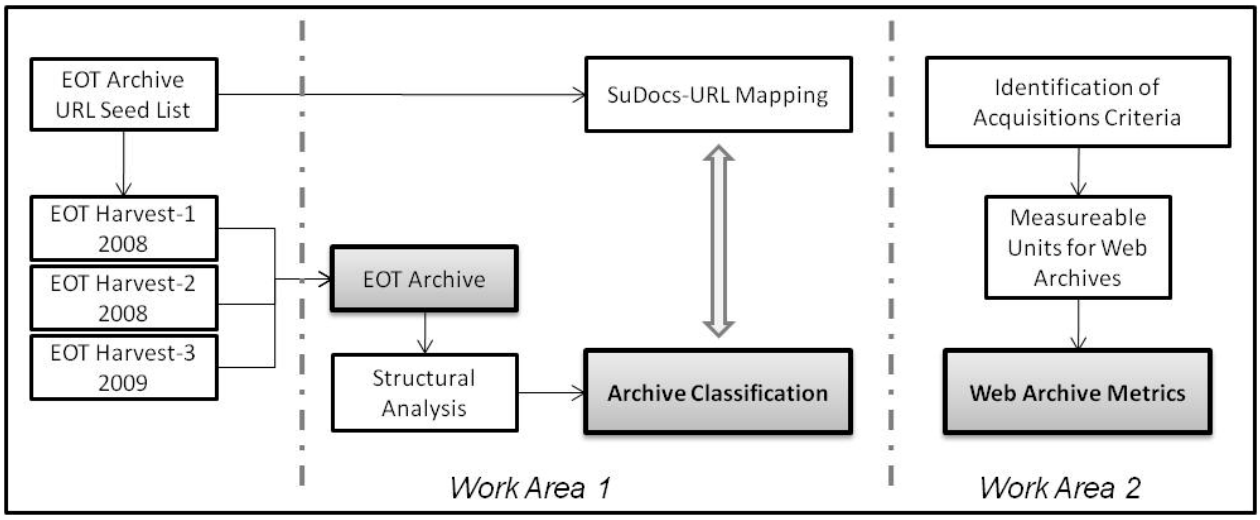


Figure 1. Project Work Areas

I. Interim Goals Accomplished

A. Archive Classification

1. Structural Analysis of Archive
   - Optimized output
   - Applied link analysis tool & created web graphs
   - Created multiple visualizations of the web graphs
2. Classification Tool
   - Developed Web-based tool for classification of Archive's URLs by SMEs
   - QA of classification tool
   - Trained SMEs
   - SMEs classified EOT Archive URLs by SuDocs scheme

B. Web Archive Metrics

1. Identification of Acquisitions Criteria
   - Further analysis of library statistics and measurements: ARL & COUNTER

- Analyzed content of focus group discussion and published findings

2. Survey of Federal Depository Libraries
   - Survey questionnaire created and validated by SMEs
   - Survey conducted; data gathered
   - Survey responses analyzed
   - Report of survey findings published

II. Significant Findings, Lessons Learned, & Accomplishments

Archive Classification

Structural Analysis of Archive

Archive Size Management

As stated in the June interim report, due to the extremely large size of the EOT Archive (Total URLs = 160,156,233), a decision was made to limit the classification scope to two domains: .gov and .mil. Together they include 141,334,979 URLs and 16,015 unique sub-domains. This number of sub-domains remained too large for both: (a) effective visualization of the underlying web graphs resulting from the link analysis and (b) feasible human classification effort on the part of the project's ten SMEs.

The URLs were converted to SURT formats[1] to evaluate possibilities for reducing their number based on unique sub-domains within the domain structures (Table 1). A decision was made to limit the structural analysis of the Archive to unique second-level domains, which resulted in 1,151 URLs for the subsequent analyses. This number should be adequate for evaluating the effectiveness of the structural analysis.

URL	http://marriagecalculator.acf.hhs.gov/marriage/			
SURT form	http://(gov,hhs,acf,marriagecalculator,)			
Domain structure	gov	hhs	acf	marriagecalculator
	Domain	Sub-domain 1	Sub-domain 2	Sub-domain 3
	1st level	2nd level	3rd level	4th level

Table 1. SURT Domain Structure Example

Archive Visualization

A number of visualization tools were investigated to depict the link relationships. Typical results are available on the project wiki – Link Analysis[2]. Many of the graphs have are interactive and can be manipulated, for example to enlarge the images or to rearrange the visualizations.

- GUESS Visualizations of NIH (National Institute of Health)
  - We started looking at the inlinks and outlinks of nih.gov because a list of NIH's family of websites was available. Once the graphs were generated, we were then able to examine the link relationships of these second-level domains with NIH. From examining these inlinks and outlinks, we have learned that looking solely at the number of links between two second-level domains without context is not enough to reliably inform a

[1] Essentially, the SURT form of the URLs inverted the order of the dot-separated fields in the domain structure.
[2] http://research.library.unt.edu/eotcd/wiki/Category:Link_Analysis

relationship due to the varying design of websites and the size of organization divisions, among other factors. It was decided to consider the ratio of total links to/from a second-level domain vs. number of links to/from a specific second-level domain in future visualizations.

- Hypergraph Visualizations of NIH
  - Visualizing more relationships than those representing NIH links became too difficult to read. The resulting images looked like rubber band balls because of the high number of links between nodes.
- Clustered Graph Visualizations
  - Produced cluster charts of the second-level domains. The charts indicate strong correlations among domain subsets.
- Treemap Visualizaions of GPO (Government Printing Office)
  - Interactive treemaps relating to GPO subdomains and mimetypes were created and provide another visualization of the Archive's structure and contents.
- Protovis Force Directed Visualizations[3] of HHS (Health and Human Services)
  - Created visualizations of eight HHS known sub-agencies and the second-level domains to which over 1% of their outlinks point (Figure 2). The eight sub-agency nodes are colored uniquely. The width of the links is directly related to the percentage of outlinks pointing to the target nodes. The size of the nodes is based on the number of other visible nodes that have links to them, although in cases where edges of two visible nodes has a weight less than the requisite 1%, these edges are not visible.
  - Created visualization of all nodes and links across the set of 1,151 .gov and .mil second-level domains where edge weights are at or above 20% of the total outlinks for a node. (Note: There are 48,000 edges in total.)

The resulting visualizations show promise in identifying clusters of related websites within the Archive. Future work will compare the sites within these clusters to the SuDOC classification assignments of the SMEs. Some future avenues of exploration have emerged as a result of the link analysis and subsequent visualizations. In the case of the EOT Archive and the SuDOC classification scheme: (a) Can we identify additional URLs in the visualizations that are associated with an agency author or a cluster group but are not classified as such, and (b) Can we account for URLs that are classified for particular agency authors but do not appear in the visualization as associated with that agency? And in the general case, is it possible to characterize certain types of sites (e.g., portals) so that they can be predictably identified within a Web archive?

[3] Protovis visualizations will not work in Internet Explorer.

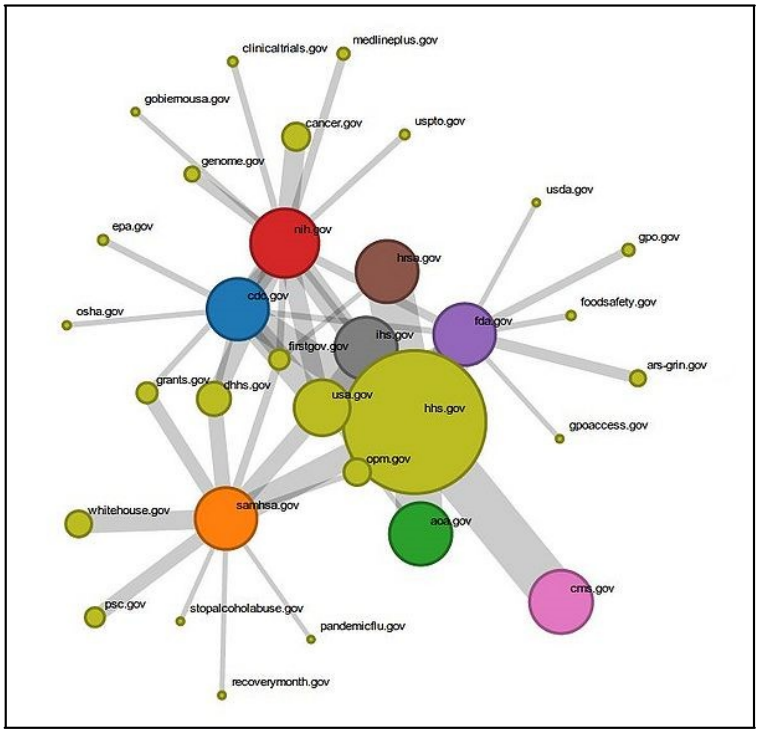


Figure 2. Protovis Force Directed Visualization of HHS[4]

Classification Tool

Requirements were specified for a web-based tool to allow the ten SMEs to classify the same set of URLs (*N* = 1,151) in the archive as are being investigated in the structural analysis. The tool was developed in Django[5], and allowed SMEs to view Websites, assign one or more SuDOCs classes to a site, and add any additional explanatory notes. SMEs could also designate sites as outside the scope of the federal government or within the scope but lacking an author listing within the SuDOC scheme.

SME Classification of Archive

The URLs were randomly assigned to the 10 SMEs for classification. To measure inter-rater reliability, each of the 1,151 was classified by two SMEs. This resulted in each SME classifying approximately 230 sites. A hands-on training session introduced the classification tool to the SMEs. The classification exercise was completed in November 2010. Future work will involve measuring the inter-rater reliability of the classification and resolving discrepancies as necessary and possible.

Web Archive Metrics

Identification of Acquisitions Criteria

[4] An interactive view of this visualization may be visited at http://research.library.unt.edu/visualization/force/hhs_agency.html. A view with arrows and link labels is at http://research.library.unt.edu/visualization/force/hhs_agency_labels.html. Hovering over the edges indicates the direction of the edge.

[5] Django is an open source, high-level Python Web framework.

Material/Resource Selection Criteria

Findings from the metrics focus group[6] identified the following selection criteria of importance to the project's SMEs. While the discussion was primarily concerned with electronic resources, there is little doubt that SMEs anticipate materials in Web archives will be akin to electronic resources in terms of the selection and acquisition decisions libraries will make. Web archive service providers will likely need to furnish information that allows libraries to evaluate archived content along the following dimensions:

- Broadness of applicability
- Usage data
    - Generally vendor provided
    - Vendor compliance with standards needed
- Appropriateness for collection
- Number of titles
- Unique content
- Duplicate content

An important finding in regard to further work in the area of metrics for Web archives is the identification of two *essential requirements* for selection decisions:

1. Standard data elements for comparable material types
2. For networked electronic resources, counts based on IP addresses for:
    a. Specific pages and collections accessed
    b. Specific files/materials retrieved

Web Archive Services & Usage Statistics

In addition to their preservation service, Web archive service providers will have the opportunity to provide two additional services for libraries: a hosting/access service and an acquisition service (Figure 3). Findings from the focus group suggest that some libraries will want to acquire materials from an archive, in particular materials that augment the comprehensiveness of a unique collection or materials that are critical to the research focus of academicians. However, access services will be the norm for most libraries, illustrating the need for archives to be positioned to provide standardized usage data.

Libraries increasingly need to demonstrate the value and impact of their services and to optimize utilization of their resources. Usage data is critical to measuring value and impact. In this regard, there are two standards efforts of particular interest and applicability to Web archive metrics:

1. COUNTER Codes of Practice and the Standardized Usage Harvesting Initiative (SUSHI): ANSI/NISO Z39.93-2007, and
2. ISO TC46/SC8/WG9: Statistics and quality issues for web archiving.

[6] Available: http://research.library.unt.edu/eotcd/w/images/0/0f/eotcd_metrics_fg_final_rpt_krm_12augl2010.pdf

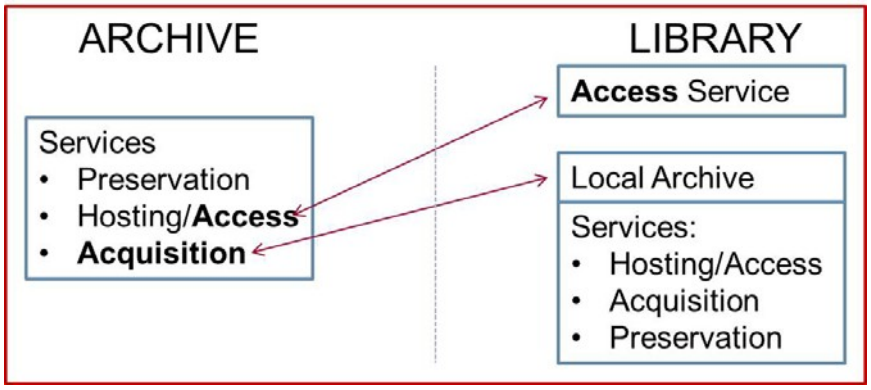


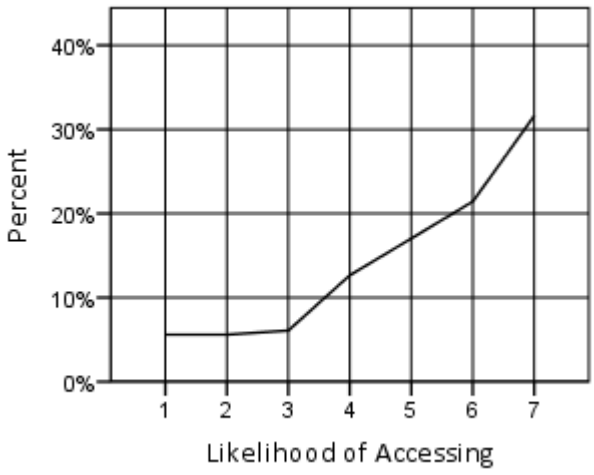
Figure 3. Web Archive Services

Survey of Federal Depository Libraries

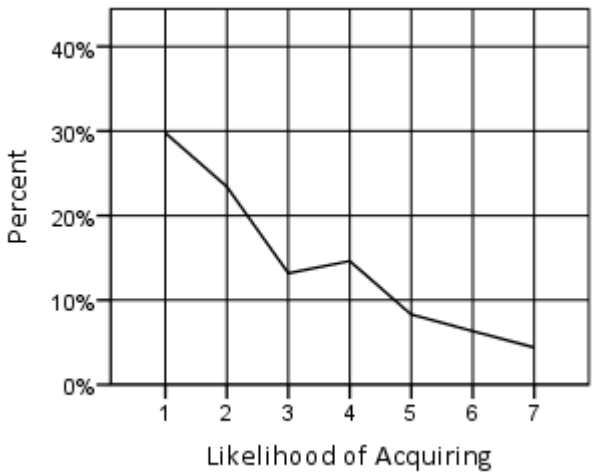
A brief online survey was conducted to assess libraries' interests in acquiring versus accessing materials in Web archives, as well as to estimate their capability to support acquisition services, such as preservation, hosting, and user access. Additionally, the relationships between three demographic characteristics (depository type, library type, and library size) and libraries' interests and capabilities were measured. The survey was conducted from September 14, 2010 – October 1, 2010. It was sent to 1225 Federal Depository libraries and a total of 414 libraries (34%) submitted responses.

The survey results[7] confirmed what the metrics focus group findings had suggested: Libraries are decidedly more likely to access materials (*Mdn* = 6) than to acquire materials (*Mdn* = 2) from Web archives at trusted institutions (Figure 4). Importantly, libraries have limited capabilities for either long-term preservation of materials acquired from Web archives or for hosting materials for user access (*Mdn* =2; range: 1 = not capable and 7 = extremely capable).

[7] Available: http://research.library.unt.edu/eotcd/w/images/2/29/fdlp_survey_report_krm_14dec2010.pdf



(*N* = 411; *Mdn* = 6)

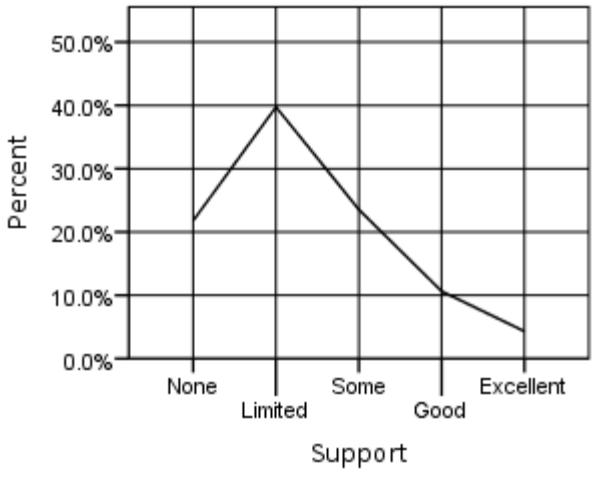


(*N* = 410; *Mdn* = 2)

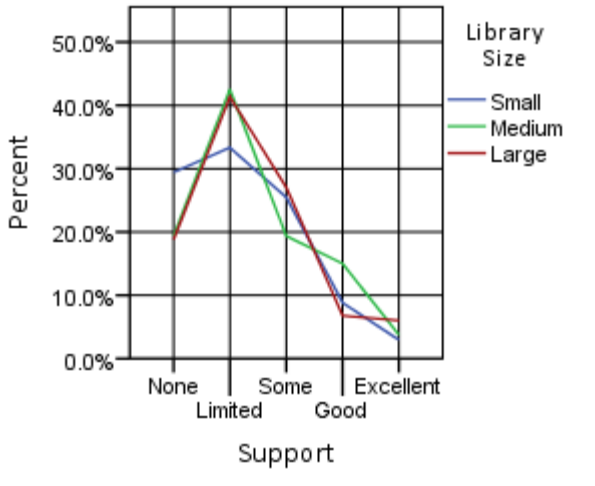
1 = Extremely unlikely, 7 = Extremely likely

Figure 4. Likelihood of Accessing versus Acquiring Materials from Web Archives

Libraries' preferences for accessing Web archives is reinforced by their estimates of the support they are likely to receive within their organizations for acquiring materials from Web archives (Figure 5). Just over 60% of libraries of all sizes (*N* = 395) indicated they had either no support (*n* = 86; 22%) or limited support (*n* = 157; 40%) for the acquisition of materials from Web archives (Figure 4). Twenty-four percent (*n* = 93) had some support, while 11% (*n* = 42) had good support and only 4% (*n*= 17) enjoyed excellent support. Not surprisingly, the likelihood of libraries to acquire materials from a Web archive was highly correlated to the support they had for acquisition within their organizations.



All Libraries (*N* = 395; *Mdn* = 2)



Libraries by Size (*N* = 392; *Mdn* = 2)

Figure 5. Support for Acquisition of Materials from Web Archives (*N* = 395; *Mdn* = 2)

III. Project Achievements

1. Findings & Reports[8]
    a. Link Analysis Visualizations
        i. Web Graph Visualizations (GUESS Visualizations, HyperGraphs, Cluster Graphs) http://research.library.unt.edu/eotcd/wiki/Web_Graph_Visualization
        ii. Web Graph Force Directed http://research.library.unt.edu/eotcd/wiki/Web_Graph_Force_Directed
        iii. Web Graph Treemaps http://research.library.unt.edu/eotcd/wiki/Web_Graph_Treemaps

    b. Metrics Focus Group Report http://research.library.unt.edu/eotcd/w/images/0/0f/eotcd_metrics_fg_final_rpt_krm_12augl2010.pdf
    c. Findings of the Web Archive Survey of Federal Depository Libraries http://research.library.unt.edu/eotcd/w/images/2/29/fdlp_survey_report_krm_14dec2010.pdf

2. Presentations
    a. Murray, K. R., Phillips, M., & Hartman, C. N. (2010, October 17). *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives.* Presented at the SME Meeting in Washington DC. Available: http://research.library.unt.edu/eotcd/w/images/0/07/Sme_mtg_dc_17oct2010.pdf
    b. Hartman, C. N. (2010, October 18). *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives.* Presented at the Federal Depository Library Conference in Washington DC. Available: http://research.library.unt.edu/eotcd/w/images/e/e9/FDLP_Conference_20101018_WashDC.pdf

3. Advisory Board
    a. Meeting with the board was held July 23, 2010 in Washington DC. In attendance: Cathy Hartman and Mark Phillips, UNT; Abbie Grotke and Gina Jones, Library of Congress; Tracy Seneca, California Digital Library; Kris Carpenter, Internet Archive; and Gildas Ilien, from the National Library of France and Chair of the ISO Committee studying metrics for Web Archives (ISO TC46/SC8/WG9). Discussion included an update on the *eotcd* project and the findings from the metrics focus group and their impact on the ISO Committee's report.

4. Subject Matter Experts
    a. Second meeting was held on October 17, 2010 in Washington, DC. All 10 SMEs were in attendance.

[8] Available on project wiki: http://research.library.unt.edu/eotcd/wiki/Main_Page