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 2011

Submitted by:

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

Kathleen Murray
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

I. Introduction

This is the third 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 December 1, 2010 – November 30, 2011.

As initially planned, the project was comprised of two work areas: (1) Archive Classification and (2) Web Archive Metrics. A no-cost extension for the project was granted for the period December 1, 2011 through November 30, 2012. Two additional areas of work are planned for this time period: (3) Improving Access to the EOT Archive and (4) Researcher Needs Assessment. The work conducted in each of the four areas is briefly described in the remainder of this section. Following the Introduction, this report includes three sections: Goals Accomplished; Significant Findings and Accomplishments; and Project Achievements.

Work Area 1 - Archive Classification

Classification of the EOT Archive involved structural analysis and human analysis. Link analysis, cluster analysis, and visualization techniques identified the organizational and relational structure of the EOT Archive and produced clusters of related websites from a representative set of the Archive's URLs. The project's subject matter experts (SMEs) classified the same set of URLs according to the SuDocs Classification Scheme using a Web-based application developed by project staff. The resulting classification served as the standard against which the effectiveness of the structural analysis was evaluated. As an additional exercise to test the topical relatedness of the clusters' members (i.e., Websites), a tool was developed to allow the project's SMEs to add subject tags to each cluster.

Work Area 2 - Web Archive Metrics

Identification of metrics for Web archives was informed by the project's SMEs who participated in two focus groups to identify and refine the criteria libraries use for acquisition decisions. A review of existing statistics and measurements used by academic libraries was conducted. Additionally, content categories for the Archive were identified. A proposed set of metrics for Web archives was created. The proposal was provided to the chair of the ISO working group (ISO TC46 SC8 WG9) that is writing a technical report, *Statistics and Quality Issues for Web Archiving*, and the PI met twice with the working group chair to discuss the proposal. Anticipating researchers' needs to understand the scope and type of content in the Archive, data elements that could be readily extracted from the Archive's files were investigated.

Work Area 3 – Improving Access to the EOT Archive

Servers will be acquired to enable experiments that integrate new functionality into existing digital library access tools. New functionality will directly relate to the integration of knowledge acquired from the cluster analysis, findings from the classification exercise, and results of the investigation of available data elements.

Work Area 4 – Researcher Needs Assessment

Interviews will be conducted with researchers, most likely in the areas of political science and environmental policy, to determine the type and range of research questions they study and to identify how the materials in the EOT Archive would assist them in their investigations. The findings from these interviews will inform a set of anticipated use cases describing how researchers' needs could be addressed.

II. Goals Accomplished

1. Archive Classification

1.1. Structural Analysis of Archive

- Completed the cluster analysis of the representative set of EOT Archive URLs

1.2. Mapping URLs to the SuDocs Classification Numbering System

- SMEs assigned SuDoc classes to the representative set of EOT Archive URLs

1.3. Classification of Clusters

- Clusters resulting from the structural analysis (1.1) were evaluated for relatedness as measured by the SuDoc classes assigned by the SMEs (1.2)

1.4. Topical Evaluation of Clusters

- A Web-based tag tool was developed for SMEs to assign subject keywords to the clusters
- Online SME tag tool training materials were created
- Analysis of the topical evaluation data was completed

1.5. Evaluation of Work Area 1

- Analysis of the effectiveness of structural analysis was completed
- Findings were presented to SMEs and Advisory Board members

2. Web Archive Metrics

2.1. Determination of Web Archive Measurement Units

- Analyzed the Archive's mime-types and identified content categories
- Created treemap visualizations of counts and sizes for the proposed content categories
- Created a proposal for Web archive metrics

2.2. Investigation of Collection Description Attributes

- Identified the core set of data elements available for the Archive's content
- Created collections in the "cdxdatabase" in MongoDB for the Archives's URIs and for the organizations that harvest the EOT Archive's content
- Created time series visualizations of the harvesting activities of the organizations

2.3. Evaluation of Work Area 2

- Presented findings and conducted a group discussion with project SMEs

III. Significant Findings & Accomplishments

Archive Classification

Structural Analysis of Archive

Due to the enormous size of the EOT Archive (Total URLs = 160,156,233), a decision was made to limit the structural analysis to unique second-level domains, which included 1,151 URLs. The following cluster analysis methods were investigated to create clusters for this set of URLs.

LinLog Clustering: Two sets of clusters

- *Set 1: 20 clusters*. The first set of clusters resulted from running the LinLog algorithm on the edges when the source and target were both in our EOTCD collection. In this case, weights were calculated as the ratio of outlinks from a source to a specific target over all outlinks from that source.

- *Set 2: 18 clusters*. As in the first set of clusters, the second set of clusters resulted from running the LinLog algorithm on the edges when the source and target were both in our EOTCD collection. In this case the weights on edges are the actual number of occurrences of a link between source and target.
- *Observations*. Using the LinLog method, we end up with some clusters that are larger than perhaps expected. We would have liked to see more clusters breaking out from these large groups. We ended up with less than half the number of clusters we hoped for based on the number of top level government author agencies.

Linlog Coordinates with Agglomerative Hierarchical Clustering: Two sets of clusters

- In this case, Linlog layout's force-directed layout techniques for weighted graphs were used to map our Web graph to Euclidean space. We then determined clusters using the agglomerative hierarchical clustering algorithm and Euclidean distance. As most popular clustering algorithms make use of Euclidean distance for their distance measure, this allowed us to create clusters based on distance in a geometric space. Two sets of clusters were produced using this method. They differ in the number of clusters defined for the algorithm.
  - *Set 1: 55 Clusters*.
  - *Set 2: 75 Clusters*.
- *Observations*.
  - We found that clustering in geometric space can be problematic when the Web graph is highly linked and its density is highly varied throughout. Laying out such a graph gives varied shapes and distances from what we would like to see as our centroids. In the EOTCD data, trying to achieve clusters that might each be representative of a single SuDoc author agency is difficult because the size of those agencies, the number and size of their subordinate agencies, and the amount that they publish differs widely.
  - However, this was perhaps our most successful clustering method and these clusters were selected for evaluating the effectiveness of the structural analysis.

Normalized Google Distance (NGD)

- In this method, we leveraged the normalized Google distance measure. While this is actually a semantic similarity measure, we have found that it translates well to our study of link analysis. In our application of this formula we measure the distance between government domains based on the similarity of their outlinks.
- Only preliminary work was conducted with this method, which resulted in a set of 76 clusters.

Strongest Outlinks and Majority Inlinks

- In this method, our starting point was our weighted Web graph where the weights were the ratio of the source's outlinks to a target over its total outlinks. The Web graph excludes links with weights less than 1%.
- This method resulted in 139 clusters that appear to be well-related.
- *Observations.*
  - By initializing with strongest outlinked clusters, we have unfortunately already eliminated 13 author agencies as centroids.
  - Because we don't have outlink data for 16 sites, they were removed from the cluster calculations.

Web Communities

- Once again, in this method our starting point was the weighted Web graph where the weights are the ratio of the source's outlinks to a target over its total outlinks. The Web graph excludes links with weights less than 1%.
- As with the NGD method, only preliminary work was conducted with this method, which resulted in 122 clusters.

Mapping URLs to the SuDocs Classification Numbering System

The SMEs completed classification of the 1,151 URLs from the EOT Archive in November 2010. Each of the URLs was classified by two SMEs. In 70% of cases, the two SMEs' classifications were in agreement (*n* = 808). In 30% of cases, the two SME's classifications were in disagreement (*n* = 343). Three arbitrators, who were experts in the SuDocs Classification Scheme, evaluated these URLs and resolved the disagreements.

Overall, the SMEs thought the SuDocs Classification Scheme worked well to classify the websites. They assigned SuDoc classes to 1,040 sites and identified a need for new SuDoc classes for 60 sites. (The remaining 51 sites were determined to be outside the scope of the federal government's domain.)

However, they agreed that the SuDocs Classification Scheme lacks sufficient granularity for subordinate offices and agencies. Oftentimes, they were forced to classify at a high level within the hierarchical SuDocs scheme, which associates classification numbers with parent agency authors within the federal government as well as the subordinate agency authors of each parent. The major challenges the SMEs experienced were: (a) determining a primary author among several authors listed on a website; and (b) discovering the actual content author on sites served by a separate hosting agency.

Classification of Clusters

The SuDoc authors determined by the SMEs and arbitrators were mapped to the members of the two cluster sets resulting from the Agglomerative Hierarchical Clustering method: 55 clusters and 75 clusters. Because SuDocs is a hierarchical numbering scheme that includes a unique alpha code for each agency, it was possible to determine the number of parent agency authors assigned to each of the clusters (Figure 1).

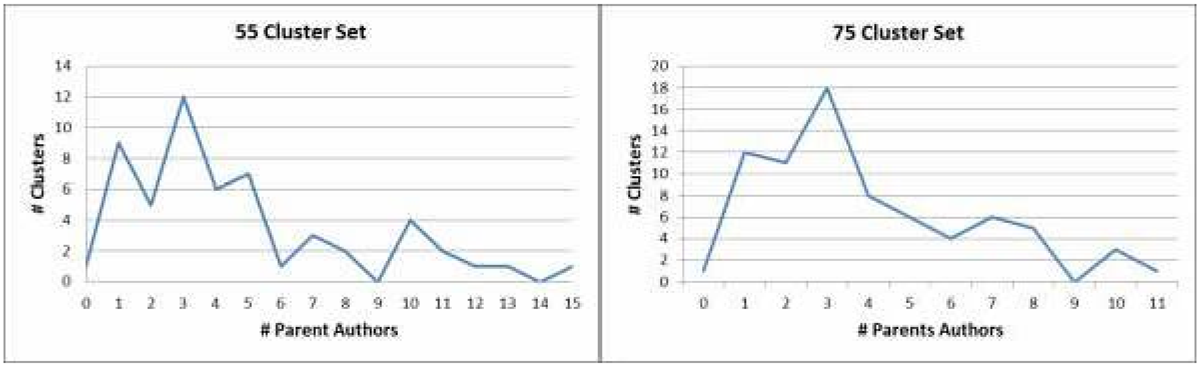

Figure 1. Number of parent authors in cluster sets

We found that increasing the number of clusters from 55 to 75 resulted in more clusters having fewer parent authors. For example, nine clusters in the 55-set had only one parent author, while 12 clusters in the 75-set had only one parent author. This was also reflected in the percentage of clusters with two or fewer parents and with four or fewer parents (Table 1).

Set of 55 Clusters		Set of 75 Clusters	
# Parents	% Clusters	# Parents	% Clusters
≤ 2	27%	≤ 2	32%
≤ 4	60%	≤ 4	67%
≤ 15	100%	≤ 11	100%

Table 1. Percentage of clusters by number of SuDoc parent authors

Topical Evaluation of Clusters

Subsequent to the classification of the clusters, we wondered if clusters with multiple SuDoc parent authors might represent topically related content from the websites of different government agencies. A tag tool was developed to allow 12 SMEs to evaluate the two sets of clusters (*N* = 130) and assign keywords and/or Library of Congress Subject Headings to each cluster. All clusters were evaluated by three SMEs. Content analysis of the tags resulted in each cluster being assigned a relatedness category (RC): 1 = little or no relation; 2 = somewhat related; or 3 = strongly related.

The findings indicate that the cluster analysis successfully identified strongly related content in 61% of clusters. There was extremely little variance in the percentage of clusters in each of the three relatedness categories among the 55-set, the 75-set, and the combined set (Table 2).

Clusters	RC 1	RC 2	RC 3
130	21%	18%	61%
75-Set	21%	17%	61%
55-Set	20%	20%	60%

Table 2. Percentages of clusters by relatedness category (RC)

Table 3 identifies the average relatedness score for three groups of clusters in the 75-set. Each group accounts for approximately one-third of the 75 clusters. Groups 1 and 2 have the fewest number of parent authors and are substantially more topically related than the clusters in group 3. It appears that the clustering method was useful in identifying topically related content across a small number of different parent agency websites. This finding may be useful in suggesting relevant content to users of future EOT Archive search systems.

Group	# Parents	% Clusters in 75-set	Average Relatedness Category *
1	≤ 2	32%	2.76
2	3-4	35%	2.65
3	5-11	33%	1.69

* 1: little or no relation; 2: somewhat related; 3: strongly related

Table 3. Average relatedness category for clusters based on number of SuDoc parent authors

There were 39 identical clusters in the 55-set and the 75-set. Seventy-two percent (*n* = 28) of these clusters had strongly related content (Table 4; RC3). The 16 remaining clusters in the 55-set subdivided into 36 clusters in the 75-set. A higher percentage of these 36 clusters were in RC3 (64%) than were the 16 clusters in the 55-set (44%) from which they derived.

# Clusters	Cluster Set	Relatedness Category *		
		RC 1	RC 2	RC 3
130	Combined sets	21%	18%	61%
39	Identical in both sets	18%	10%	72%
16	Unique to 55-Set	25%	31%	44%
36	Unique to 75-Set	22%	14%	64%

* 1: little or no relation; 2: somewhat related; 3: strongly related

Table 4. Average relatedness category for clusters based on number of SuDocs parent authors

We found that specifying a larger number of clusters in the cluster analysis algorithm resulted in more clusters whose members' websites contained content that was strongly related. While the optimal number of clusters to specify is an unknown, it is helpful to know that more topically related content is likely to be identified by specifying larger numbers. In our project this translates to numbers greater than the number of actual parent agencies in the SuDocs scheme. Additionally, clusters that contain the websites of a single federal government parent agency are more likely to be identified by specifying larger numbers.

Further analysis of the 75 cluster set was done to identify whether the numbers of cluster members, total SuDocs authors (i.e., both parent and subordinate agencies), or only SuDocs parent authors impacted the clusters' relatedness categories. As illustrated in Table 5, neither the average numbers nor the ranges for these three characteristics varied substantially across the relatedness categories. However, there was a decreasing trend in the average number of SuDoc parents as the relatedness of the clusters increased. This is consistent with the data reported in Table 3.

Cluster Set Characteristics	Relatedness Category *		
	1	2	3
# Clusters (*N* = 75)	*n* = 16	*n* = 13	*n* = 46
# Cluster Members			
average	15	12	16
range	3-48	3-30	2-53
# SuDoc Authors			
average	8	6	6
range	2-16	2-14	0-15
# SuDoc Parents			
average	6	4	3
range	2-11	1-8	0-9

* 1: little or no relation; 2: somewhat related; 3: strongly related

Table 5. Averages and ranges for characteristics of the 75 cluster set by relatedness category

Evaluation of Structural Analysis

As noted previously, we found that the Linlog Coordinates with Agglomerative Hierarchical Clustering method produced the best results among the five clustering methods investigated. The results of the SuDoc classification exercise, which involved human subject matter experts, indicated that in 67% of the clusters in the 75-set and in 60% of clusters in the 55-set the structural analysis was effective at creating clusters of related websites created by four or fewer SuDocs parent authors (Table 1). Both the classification exercise and the subject tagging exercise indicated that increasing the number of clusters specified in this clustering method resulted in: (a) more clusters with fewer SuDocs parent authors and

(b) more topically related clusters.

Figure 2 illustrates the percentage of clusters in the 75-set by relatedness category and the number of SuDoc parent authors. This figure is another view of the effectiveness of the structural analysis, indicating that the highest percentages of clusters containing websites with either strongly related content (RC3) or somewhat related (RC2) content had four or fewer SuDoc parent authors. Conversely, the highest percentages of clusters whose content had little or no relationship (RC1) had greater than four parent authors.

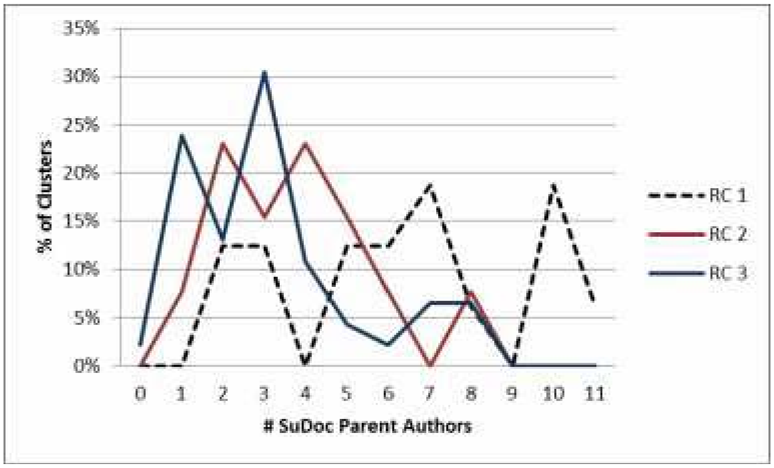

Figure 2. Percentage of clusters by relatedness category and number of parent authors (N = 75)

Web Archive Metrics

Determination of Web Archive Measurement Units

In light of the findings of the initial focus group discussion with the project's SMEs, and after an analysis of the statistics reported by academic libraries, it was determined that the ARL Supplementary Statistics categories for the *Use of Networked Electronic Resources & Services* and for *Library Digitization Activities* were key existing measures to evaluate for their possible application to Web archive metrics. Because some the statistical categories in regard to the use of databases and services specify data derived from the *COUNTER Code of Practice*, that specification was also evaluated for its application to Web archive metrics.

We determined that, in general, there are four categories of measurement for which academic libraries collect data:

1. Scope (How much; how many)
2. Expenditures (Cost)
3. Usage (Counts)
4. Quality (Outcomes; Value)

Of these, the project's SMEs identified two critical areas for which Web archive statistics will be needed to inform their selection and retention decisions: Scope and Usage. These two areas were the primary focus of our metrics proposal.

Web Archive Metrics Proposal

SCOPE

An objective of this project is to suggest metrics that characterize the resources in a Web archive in a manner that is meaningful to librarians and library administrators, who range in their degree of familiarity with the technical definitions employed by standards bodies and the wider technical community.

To meet this objective, we analyzed the content of the EOT Archive by mime types and subsequently identified categories for some of the resource formats associated with the "application" and "text" mime types. The resulting content categories are listed in Table 6. The categories suggest aggregate measurement units for Web archive resources. Treemap visualizations of the sizes and counts within the EOT Archive for the proposed content categories were produced.

Category	# URIs	# Formats	Formats
text	109,498,363	2	html, plain
image	29,140,868	8	jpeg, gif, png, tiff, pjpeg, x-icon, jpg, bmp
document-like	11,234,522	4	pdf, msword, postscript, vnd.ms-powerpoint
computer files			
* coded/formatted	2,427,349	11	x-javascript, javascript (both text and application type), x-cgi, xml (both text and application type), atom+xml, rss+xml, x-vcal, x-vcalendar, css
* compressed	526,105	5	zip, x-zip-compressed, x-gzip, x-compress, vnd.google-earth.kmz
* binary	503,660	2	octet-stream, x-octet-stream
* executable	15,079	1	download
dataset	908,339	5	vnd.ms-excel, csv, comma-separated-values, x-netcdf, fits
video	318,498	5	quicktime, x-ms-asf, mpeg, x-ms-wmv, x-shockwave-flash
audio	198,349	3	mpeg, x-pn-realaudio, x-wav

Table 6. Content categories within the EOT Archive

PROPOSED DATA ELEMENTS for SCOPE

1. For a Web archive:
    a. Size (in gigabytes, terabytes, etc. as appropriate)
    b. Number of discrete collections
2. For each collection within a Web archive:
    a. Size (in gigabytes, terabytes, etc. as appropriate)
    b. Number of objects by type:
        i. Text
        ii. Image
        iii. Document-like

iv. Computer file
v. Dataset
vi. Video
vii. Audio

USAGE

As mentioned earlier, in terms of statistics tracked and reported by academic libraries, Web archives most closely resemble statistics reported using the ARL supplemental statistics worksheet for the use of networked electronic resources and services. ARL includes three usage measures for databases and services and instructs libraries to derive the values for these numbers from reports specified in the COUNTER Code of Practice (Table 7).

Statistic	COUNTER Code of Practice
number of sessions	Database Reports 1 and 3
number of searches	Database Reports 1 and 3
number of successful article requests	Journal Report 1

Table 7. ARL statistics and corresponding COUNTER report

The PIRUS and PIRUS2 projects are investigating the adaptation of COUNTER usage measurements and reports for materials in institutional repositories. These investigations have a similar purpose to our investigation into usage statistics for Web archives. It seems prudent that our work to establish usage statistics for Web archives should also be informed by the COUNTER Code of Practice. It is hoped that doing so will enable libraries to evaluate their patrons' use of the materials in Web archives in the manner they are already familiar with for other classes of electronic resources (i.e., ebooks, databases, and journals).

PROPOSED DATA ELEMENTS for USAGE

1. For each collection within a Web archive:
   a. Number of sessions
      i. Total number
      ii. Number federated or automated
   b. Number of searches (queries)
      i. Total number of searches run
      ii. Number federated or automated

Investigation of Collection Description Attributes

Perspectives on Content Description for Web Archives

User Perspective

We were concerned with one class of user, a library. We asked librarians serving as project SMEs what criteria their libraries used in making acquisition decisions. From their responses we discovered that describing an archive's content is essential and goes beyond measures of its scope. Further, libraries require consistency in content descriptions for the same type of materials that are available from different providers.

Content description allows a library to assess the broadness of applicability of all, or a portion of, a provider's content to a library's collection. For libraries, this assessment is fundamental in their material selection process. We identified three attributes to consistently describe a collection within a Web archive:

1. Topical areas covered
2. Unique or exclusive content available
3. Dates materials were harvested

Provider Perspective

Content description is important to Web archive providers for a few reasons: (a) to determine change-over-time for similar content captured at different points in time; and (b) to identify content overlap among collections. It seems reasonable that, if reported in a consistent manner, these characteristics of a Web archive will promote access and discovery of materials.

Common Attributes

The two perspectives share common attributes for content description. We suggest the following:

- Topical areas addressed
  - At a feasible level of effort, whether resulting from human mediation or machine analysis
- Unique or exclusive content available
  - Dates materials in the collection were captured
  - Measure of how the collection changed-over-time
  - Analysis of collection's overlap with other known collections

Core Data Elements Available

One statistic ARL requires libraries to report for database usage is the “number of successful article requests” as reported in the vendor-provided *Journal Report 1* specified in the COUNTER Code of Practice. We did not include a corollary to this in our metrics proposal because further investigation is needed to understand how this applies to Web archives.

The COUNTER definition is the number of items requested by users as a result of a search, for example, server-controlled viewing, downloading, emailing, and printing. We recommend that use cases in this regard be developed for Web archives. We are specifically interested in understanding the core data elements within the EOT Archive’s W/ARC files that need to be extracted so that users’ discovery requirements for the search system can be accommodated. We began work in this area by (a) identifying the data elements that are currently available for the EOT Archive or that can be calculated and (b) experimenting with MongoDB, an open source, schema-free, document-oriented database.

CDX Files

The data used for the analysis of mime-types to identify content categories within the EOT Archive was extracted from CDX Files. The CDX files themselves were extracted from the Archive’s W/ARC files using extraction tools, many developed by the Internet Archive. A list of the data elements available in our CDX files is on the project wiki.

MongoDB Collections

Previously, we used Redis for storing and querying the CDX data used in this project. We began experimenting with MongoDB for several reasons including: indexing purposes, Python driver availability, and a built in map/reduce functionality. Two collections of the "cdxdatabase" in MongoDB were created: "uris" and “daily”.

The “uris” collection contains 160,000,000+ documents representing the URIs in the EOT Archive. To aggregate the various pieces of data we had for each URI, we matched up sizes of objects with the data from their CDX file records. Because URIs can occur more than once in an archive collection (i.e., if the URI is crawled multiple times by multiple institutions), we looked at the time stamp, the W/ARC the URI instance came from, and the checksum. Additionally, we calculated other information, including: the SURT form of the URI, the Domain SURT form for the URI, the harvesting organization the URI should be attributed to, and the top level domain. Currently we have indexes on: _id (default), time stamp, mime type, and org.

Each document in the “daily” collection contains the following data: (a) the total URIs downloaded per day and by institution, (b) total bytes downloaded per day and by institution, and (c) total URIs and bytes for items with http status of 2XX (i.e., OK) per day and by institution. From this data, time series visualizations of the harvesting activities of the organizations responsible for harvesting EOT Archive content were created. Example documents from the “daily” and “uris” collections are available on the project wiki.

Evaluation of Work Area
Our metrics proposal was provided to the chair of the ISO working group (ISO TC46/SC8/WG9) that is creating a technical report regarding metrics for Web Archives. We were given the opportunity to comment on an early draft of the report. We found there was a good deal of congruence between their technical report and our proposal and findings in regard to content description. One difference was that the technical report is more reflective of the needs for metrics at national libraries while our work is more reflective of the needs of academic libraries.

The proposed metrics for the scope and usage of Web archives, as well as the descriptive attributes for Web archive contents, were discussed with project SMEs in a focus group in October 2011. Both were endorsed by the SMEs, many of whom welcomed the incorporation of COUNTER-compliant reports. Overall there was a sentiment expressed by the SMEs that participation in this project had been educational, with many gaining an increased appreciation for the content being captured and preserved in Web archives as well as insight into the value Web archives will offer future researchers.

IV. Project Achievements

1. Papers & Reports[1]
    a. SuDoc Classifications of Clusters Resulting from Cluster Analysis Methods
    http://research.library.unt.edu/eotcd/wiki/SuDoc_Classifications_of_Clusters_Resulting_from_Cluster_Analysis_Methods
    b. Web Archive Service Models and Metrics
    http://research.library.unt.edu/eotcd/wiki/Web_Archive_Service_Models_and_Metrics
    c. Available Data About EOTCD Content
    http://research.library.unt.edu/eotcd/wiki/Available_Data
    d. Data Work
    http://research.library.unt.edu/eotcd/wiki/Data_Work

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

2. Presentations
    a. Murray, K. (2011, October). *Curation of the End-of-Term Web Archive*. Presented at the Federal Depository Library Conference, Washington, DC.
    b. Murray, K. R. (2011, October 16). *Classification of the End-of-Term Archive.* Presented at the SME Meeting in Washington, DC. Available: http://research.library.unt.edu/eotcd/w/images/3/3b/DC_2011.pdf
    c. Murray, K., Ko, L., & Phillips, M. (2011) *Curation of the End-of-Term Web Archive*. Proceedings of the Archiving Conference of the Society for Imaging Science and Technology, 8, 71-76.
    d. Murray, K. R. (2011, April 3). *Classification of the End-of-Term Archive: Status and Interim Findings.* Presented at the SME Meeting in San Antonio, TX. Available: http://research.library.unt.edu/eotcd/w/images/5/5e/Sme_mtg_sat_03apr2011_krm_07apr2011.pdf

3. Advisory Board
    a. A meeting with the board was held November 4, 2011 via Web conference. In attendance: Cathy Hartman , Mark Phillips, Lauren Ko, and Kathleen Murray, UNT; Kris Carpenter, Internet Archive; Abbie Grotke and Gina Jones, Library of Congress; and Tracy Seneca, California Digital Library. A presentation reporting the project's findings was delivered. Discussion primarily concerned ideas for future work to build on the findings from the EOTCD project.

4. Subject Matter Experts
    a. The third meeting was held April 3, 2011 in San Antonio, TX. Twelve SMEs were in attendance, including two new SMEs who had served as arbitrators in the classification exercise.
    b. The fourth and final meeting was held on October 16, 2011 in Washington, DC. Eleven SMEs were in attendance.