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

Classification of the End-of-Term Archive:

Extending Collection Development Practices to Web Archives

FINAL REPORT

FEBRUARY 2013

Submitted by:

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

Kathleen Murray
Project Coordinator
kathleen.murray@unt.edu

Mark Phillips
Co-Principal Investigator
Mark.phillips.unt.edu

Authorized Organizational Representative:

Britt Krhovjak
Assistant Director for Research Accounting
University of North Texas
UNT Libraries
1155 Union Circle #305190
Denton, TX 76203-5017

Contents

I. Introduction

This is the final report for the *EOTCD* project, which is formally titled *Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives*. The project commenced December 1, 2009 and ended November 30, 2012. The overview includes background information about the End of Term (EOT) 2008 Archive and a brief description of the activities conducted in the project's four work areas. Following the Overview there are three sections: Goals Accomplished; Significant Findings and Accomplishments; and Project Achievements.

End of Term 2008 Archive

In 2008 five United States institutions collaborated to archive the U.S. federal government Web presence: the Library of Congress, the Internet Archive, the California Digital Library, the Government Printing Office, and the University of North Texas (UNT). Their objective was to document the changes coincident with the shift in leadership of the U.S. executive branch. The resulting End of Term (EOT) 2008 Archive is comprised of 160,211,356 URI's captured during a seven month period from August 2008 to March 2009.

All harvested content was stored in either the ISO standard WARC format or the legacy ARC format (WARC/ARC). Each institution responsible for harvesting content packaged their content using the BagIt file packaging format and subsequently transferred a copy to the Library of Congress, which served as the central data collector for the Archive. The UNT Libraries acquired the dataset in the summer of 2009. The Web archiving staff at the Internet Archive consulted with the UNT programming staff regarding implementation and refinement of available archive analysis tools developed by the Internet Archive.

The EOT 2008 Archive dataset was groomed for ingestion into the UNT Libraries Digital Collections. One copy was ingested into the UNT Digital Archive, a preservation repository. A second copy was staged on public-facing servers for access. An instance of the Open Wayback Machine provided user services. In conjunction with implementing the Open Wayback Machine instance, a comprehensive CDX file containing information about the URLs present in a Web archive was created.

Analysis of the EOT2008 Archive's CDX file identified: (a) the five largest Top Level Domains based on the number of URLs and subdomains (Table 1); and (b) the top four file formats by number of mime-type (Table 2).

Top Level Domains	# URLs	# 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 URLS & subdomains by top level domains

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

Table 2. EOT Archive mime-types by number of files

As initially planned, this two-year 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 were planned for this time period: (3) Improving Access to the EOT Archive and (4) Researcher Needs Assessment.

Areas of Work

The activities of the project were carried out in four areas: Archive Classification, Web Archive Metrics, Improving Access to the EOT 2008 Archive, and Researcher Needs Assessment. The key activities in each area are described in the remainder of this section. Further details about the work conducted, as well as the findings and accomplishments are described in the sections that follow.

Work Area 1 - Archive Classification

Classification of the EOT 2008 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

The Portable Document Format (PDF) files in the EOT 2008 Archive represent a class of content many information professionals associate with the traditional notion of "discrete documents". Over four million unique PDF documents were extracted from the Archive and a series of metadata and information extraction processes were conducted for each document. Additionally, derivative raster images of the first page of each document were created. These metrics were ingested into a database for further analysis, which brought to light previously hidden characteristics of the federal government's Web-published content.

Work Area 4 – Researcher Needs Assessment

Interviews were conducted with researchers in several academic disciplines to determine the type and range of research questions they study as well as to identify how the materials in the EOT 2008 Archive might assist them in their investigations. The interviews were content analyzed to identify disciplines whose researchers might find the contents of the EOT Archive of interest, and to identify their access and discovery needs.

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

3. **Improving Access to the EOT 2008 Archive**
   - 3.1. Extraction of PDF Dataset
     - Identified PDF documents in the Archive
     - Extracted the unique PDF documents based on hash values
   - 3.2. Creation of a "PDF sample" per Document
     - Extracted data from each file:
       - Full-text of the PDF file
       - Image files of the first page (high resolution and thumbnail)
       - Embedded metadata
       - Language

  - Names, place, organizations
  - CDX file record
- Indexed dataset

3.3. Characterization of PDF dataset

- Identified queries of interest
- Characterized dataset
  - Domains and subdomains
  - Size
  - PDF format
  - Embedded metadata

3.4. Evaluation of Work Area 3

- Analysis of the PDF dataset completed

4. Researcher Needs Assessment

4.1. Interview Protocol

- Created questionnaire
- Completed IRB review

4.2. Faculty Interviews

- Identified researchers
- Conducted interviews

4.3. Data Analysis

- Transcribed interviews
- Completed content analysis

4.4. Evaluation of Work Area 4

- Analysis of interviews completed

III. Significant Findings & Accomplishments

Work Area 1 - Archive Classification

Structural Analysis of Archive: Link Analysis

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. Three cluster analysis methods were investigated to create clusters for this set of URLs: (1) LinLog Clustering, (2) Linlog Coordinates with Agglomerative Hierarchical Clustering, and (3) Strongest Outlinks and Majority Inlinks. Additionally, preliminary investigations were conducted with two other clustering methods: (4) Normalized Google Distance (NGD) and (5) Web Communities. However, limited time and resources prohibited sufficient exploration of these two methods.

After evaluating the clusters resulting from the first three analyses, a judgment was made to utilize the clusters from the Linlog Coordinates with Agglomerative Hierarchical Clustering to evaluate the effectiveness of the structural analysis. A brief summary of the results and our observations about the three methods follows.

Method 1. LinLog Clustering: Two sets of clusters

> Set 1: 20 clusters[1]. 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[2]. 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.

Method 2. 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

[1] http://research.library.unt.edu/eotcd/w/images/8/82/Linlog_clusters_ratio_weights.txt

[2] http://research.library.unt.edu/eotcd/w/images/d/d6/Linlog_clusters_not_ratio_weights.txt

space. Two sets of clusters were produced using this method: (Set 1) 55 clusters[3] and (Set 2) 75 clusters[4]. They differ in the number of clusters defined for the algorithm.

Observations

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.

Method 3. 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[5].

Observations

By initializing with the 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.

Method 4. 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[6].

Method 5. 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[7].

Archive Classification

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 =

[3] http://research.library.unt.edu/eotcd/w/images/d/d0/Clusters_linlog_agglom_euclid_55.txt

[4] http://research.library.unt.edu/eotcd/w/images/0/02/Clusters_linlog_agglom_euclid_75.txt

[5] http://research.library.unt.edu/eotcd/w/images/b/b6/Clusters_outlinks_lauren_webgraph.txt

[6] http://research.library.unt.edu/eotcd/w/images/0/08/ngd_clusters_1.txt

[7] http://research.library.unt.edu/eotcd/w/images/5/5f/Clusters_communtities_centroid_based.txt

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. [http://research.library.unt.edu/eotcd/wiki/In_Scope_-_Unable_to_Classify_List] (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: set 1 with 55 clusters and set 2 with 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).

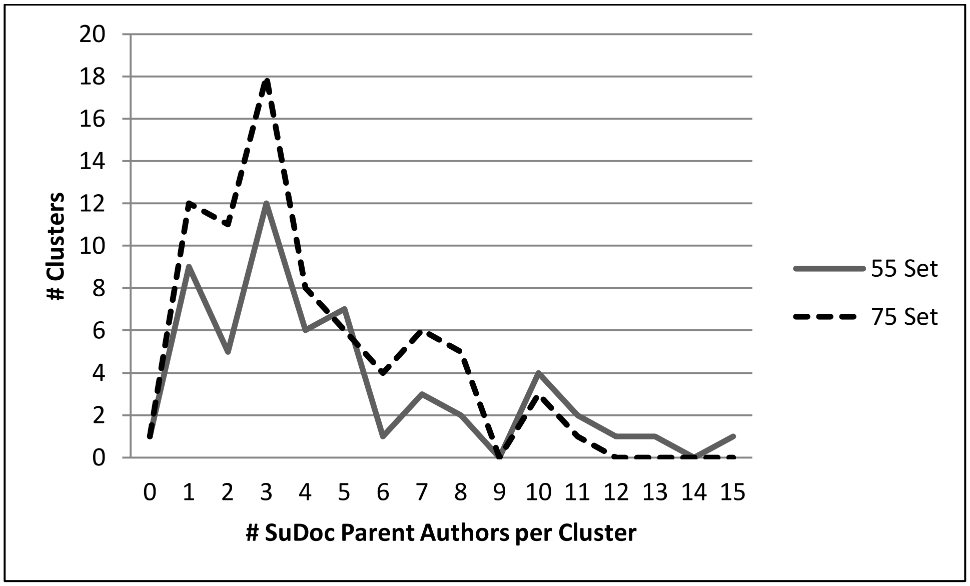

Figure 1. Number of parent authors in cluster sets

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 3).

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 3. 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 4).

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

Table 4. Percentages of clusters by relatedness category (RC)

Table 5 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 (75-Cluster 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 5. 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 6; 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 6. 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 7, 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 7. 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 3). 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.

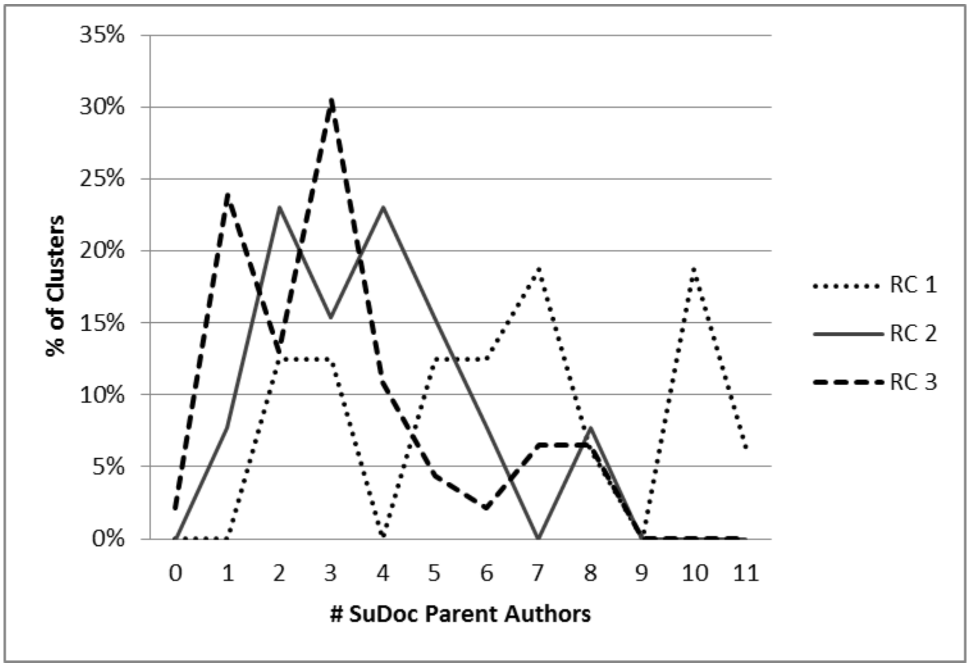

Figure 2. Percentage of clusters by relatedness category and number of parent authors (N = 75)

Work Area 2 – Web Archive Metrics

Determination of Web Archive Measurement Units

In light of the findings of the initial focus group discussion[8] with the project's SMEs, and after an analysis of the statistics reported by academic libraries, it was determined that the Association of Research Libraries (ARL) Supplementary Statistics categories[9] 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 8. 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[10].

[8] http://research.library.unt.edu/eotcd/wiki/Acquisition_Criteria
[9] http://research.library.unt.edu/eotcd/wiki/ARL_Supplementary_Statistics
[10] http://research.library.unt.edu/eotcd/visualization/treemaps/eot_metrics_treemap.html

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 8. 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 9).

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 9. 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., e-books, 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

Definitions[11]:

Session: A successful request of a Web archive service. It is one cycle of user activities that typically starts when a user connects to Web archive and ends by terminating activity that is either explicit (by leaving the service through exit or logout) or implicit (timeout due to user inactivity). Sessions can be initiated by regular searches, automated searches, or federated searches.

Search (Regular): A user-driven intellectual query, typically equated to submitting the search form of the Web archive service to the server.

[11] Adapted from COUNTER Code of Practice for e-Resources, Appendix A: Glossary of Terms. Updated 30 November 2012. Retrieved February 26, 2013 at http://www.projectcounter.org/r4/APPA.pdf

Automated Search: An automated search originates from a discovery layer or similar technology and searches multiple Web archives simultaneously with a single query from the user interface. The end user is not responsible for selecting which Web archives are being searched.

Federated Search: A federated search program allows users to search multiple Web archives owned by the same or different service providers or vendors simultaneously with a single query from a single user interface. The end user is not responsible for selecting which Web archives being searched.

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 typical CDX file entry for a URL contains nine fields separated by a whitespace character. Table 10 defines the fields in the EOT2008 Archive CDX files.

Field Name	Value
canonicalized URL	1010ez.med.va.gov/sec/vha/1010ez/form/vha-10-10ez.pdf
timestamp	20090118033012
URL	https://www.1010ez.med.va.gov/sec/vha/1010ez/Form/vha-10-10ez.pdf
content-type / mime-type	application/pdf
http status code	200
hash of file content	X65KODFIETXNBOWDTJUIAFLBQTSAMW3Q
redirect information	-
offset of record in container file	21314355
WARC/ARC filename	CDL-20090118025004-00001-dp01.warc.gz

Table 10. Fields and values in a typical CDX file entry

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[12]. Example documents from the “daily” and “uris” collections are available on the project wiki[13].

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.

[12] http://research.library.unt.edu/eotcd/visualization/timeseries/eot_timeseries_daily_compare.html
[13] http://research.library.unt.edu/eotcd/wiki/Data_Work

Work Area 3 – Improving Access to the EOT 2008 Archive

Selection of PDF Content

The UNT Libraries was interested in providing government information professionals with mechanisms to identify resources of interest for their collections within the very large, and relatively inaccessible, EOT 2008 Archive. Because of the previously documented interest of government information professionals in archived PDF documents, as well as the fact that over 10 million PDF documents are represented in the Archive, the PDF files were a logical subset of content to investigate in a systematic manner. The project team sought to improve its understanding of this important class of content.

The question directing this work area was: Is it feasible to describe the content of Web archives by format-specific features? If so, it may also be feasible to take advantage of the descriptive findings and use them to inform the development of mechanisms that aid information professionals in their collection building processes.

Extraction of PDF Dataset

The CDX file was used to identify the PDF documents in the EOT 2008 Archive. A script was written to extract the data for each URL in the CDX file containing a content-type / mime-type of “application/pdf” and an http status code of “200”. There were 10,318,073 PDF documents in the resulting list. The next step was to limit the PDF documents to unique files based on their hash values. This resulted in 4,544,465 candidate PDF documents, which were extracted to form the research dataset used in this study.

A series of information extraction routines was performed on the dataset in order to create a “PDF sample” for each candidate PDF document. Each PDF sample included a PDF document, named using its unique content hash in the format of <hash>.pdf, as well as the additional files resulting from the following processes that were run on each PDF document.

- Full-text of the PDF file extracted using the *pdftotext* utility from the xPDF library and saved as a <hash>.txt file
- Two image files were created from the first page of the PDF document. The first image was a high resolution derivative at 300 dots per inch and the second image was a thumbnail image which measured 250 pixels across the horizontal of the image. These image files were generated by using the command line utility convert, which is part of the ImageMagick image manipulation toolkit. The large image was named <hash>.jpg and the thumbnail was named <hash>.thumbnail.jpg
- Embedded metadata from the PDF file was extracted using the *pdfinfo* utility, from the xPDF library. The resulting metadata was saved as a file named <hash>.meta
- The predominant language for the PDF document was identified by feeding the extracted full-text into the Java Language-Detection library, which created an ouput file of most likely languages and probabilities for those languages. This file was named <hash>.lang
- The Stanford NER library was used to extract names, places and organizations from the extracted full-text. The library’s three-class classifier english.all.3class.distsim.crf.ser.gz was used for this process. The output of the NER process was saved as a file named <hash>.ner

- The lines of the CDX file which reference this document were extracted from the master CDX file and included in a file named <hash>.cdx

A completed “PDF sample” for this project was defined as a PDF directory, named using its unique content hash, which contained the required eight files. Here is an example:

```
3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.cdx
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.jpg
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.lang
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.meta
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.ner
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.pdf
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.thumbnail.jpg
        |___ 3UN3W3YKTPX6ZC36AHYM73CF4JISEWN4.txt
```

After the 4.5 million PDF documents were processed, they were classified as either complete or incomplete samples. The distinction was based on whether a PDF file was corrupted or not. Complete samples were not corrupt and represented 97% of the dataset (*N* = 4,544,465).

Indexing & Query Formation

Thirty-five data elements were extracted from each complete PDF sample directory and serialized as JSON files. The data elements included fields extracted from the PDF document itself as well as fields derived from the full text and other data values in the sample (Table 11). The extracted JSON data files were indexed using the Solr search system. This search system provided the ability to query the dataset using the data elements, as well as the ability to aggregate and generate statistics for various metrics of interest.

Data Fields		
Author	Modification Date	Subject
Character Count	Optimized	SURT Domains
Creation Date	Orientation	Tagged
Creator	Page Area	Text Hash
Encrypted	Page Height	Text Signature
File Size	Page Width	Title
Host Domain Count	Page Size	Unique Host Domains
Host Second Level Domain Count	Number of Pages	Unique Second Level Domains
Host Top Level Domain Count	PDF Version	Unique Top Level Domains
Host URL Count	Percent Integer	Word Count
Host URLs	Primary Language	Words Per Page
Identifier	Producer	

Table 11. Data fields extracted & indexed from the PDF samples

After indexing the dataset the Solr search system was used to aggregate and generate ad-hoc statistics across the PDF collection. The Solr system allowed researchers to construct and execute several questions formulated during the processing of the dataset. These questions required the use of PDF format-specific data as well as data that is common across all Web archive content. Example queries included: *Which domain publishes the most PDF documents? What is the average number of pages per PDF document?* The Solr system returns XML or JSON responses that were parsed and integrated into Microsoft Excel for further analysis.

Characterization of the PDF Dataset

Key findings from the analysis of query response data from the Solr system are discussed in four categories. These are: Domains and Subdomains, Size, PDF Format, and Embedded Metadata.

Domains & Subdomains

Distribution of PDF Documents

All of the EOT 2008 Archive's content, including the PDF dataset, was harvested during a seven-month period from August 2008 to March 2009. The majority of content was harvested from the following five top level domains, .gov, .mil, .edu, .us, and .org. Each PDF document in the Archive had the opportunity to be harvested a number of times, either from the same URL or as a result of being hosted on multiple top level domains in the Archive (e.g., .gov and .mil). The range of top level domains hosting an identical PDF document was 1-4.

The documents also had the possibility of being hosted on multiple top level subdomains (e.g., nasa.gov and house.gov) or lower level subdomains (e.g., jpl.nasa.gov and nlm.nih.gov). On average the PDF documents in the EOT 2008 Archive were hosted on 1.1 top level subdomains, while the range was one to twenty-five.

A single PDF instance, defined as a PDF document with the same content hash, was harvested from at least one and in many cases up to 1,763 unique URLs. The high end of this range represents content that is generated consistently even when the URL changes slightly, for example if there are session ids in the URL. The average number of URLs per PDF in the Archive was 1.2.

Distribution by Page Counts

The number of PDFs harvested from different subdomains suggests content-rich subdomains versus subdomains that host less content. The top level subdomain in the EOT 2008 Archive hosting the most PDF documents was gpo.gov (the U.S. Government Printing Office). This top level subdomain hosted 1,082,735 or 25% of the PDF samples in the Archive. This number is an aggregate of all of the lower level subdomains within gpo.gov, such as access.gpo.gov or permanent.gpo.gov.

Table 12 lists three rankings for six top level subdomains in the Archive according to: (column 2) total number of PDF documents hosted; (column 3), number of one-page PDF documents hosted; and (column 4) number of PDF documents hosted that contain 20 or more pages. (NOTE: Subdomain references at the end of the paper identify the formal agency names for the top level subdomains in Table 12.)

Rank	Total # Documents	# One-page Documents	# Documents >= 20 Pages
1	gpo.gov	gpo.gov	gpo.gov
2	usda.gov	usda.go	gao.gov
3	house.gov	house.gov	epa.gov
4	army.mil	uscis.gov	usda.gov
5	bea.gov	uscourts.gov	army.mil
6	census.gov	army.mil	noaa.gov

Table 12. Top level subdomains by PDF documents and pages

Size

Number of Pages

The PDF format is often considered the most "document like" of formats on the Web. The association of PDF files with documents introduces the concept of "pages", which allows for a direct parallel between the physical and digital worlds. The project team investigated the page count of the PDF documents in the EOT 2008 Archive in order to better understand the makeup and distribution of content.

There are a total of 60,874,402 pages represented in the 4,404,048 PDF documents in the dataset. The number of pages per PDF document ranged from many instances with only one page (n = 1,477,612; 34%), to a single instance with 17,584 pages. On average a PDF document contained 13.8 pages and PDF documents containing 1 to 14 pages accounted for 84% of the documents in the dataset. While the majority of PDF files fall at or below the average page count, there are a significant number of files that are 15 pages or more in length. Table 13 shows the distribution of documents by the range of pages per document, from 1 to over 1,001 pages.

Page Range	#	%	Cumulative %
1	1,477,612	33.55%	33.55%
2-14	2,203,216	50.03%	83.58%
15-100	616,552	14.00%	97.58%
101-1,000	104,766	2.38%	99.96%
1,001+	1,902	0.04%	100.00%

Table 13. Distribution of PDF documents by range of pages

PDF Format

Versions

The PDF format and version numbers have evolved over the past two decades from the initial 1.0 release by Adobe Systems in 1993 to ISO 32000-1:2008 “Document management -- Portable document format -- Part 1: PDF 1.7” in 2008. Now an ISO (International Organization for Standardization) standard, the PDF format continues to add new functionality while striving to maintain backwards compatibility with previous versions of the specification. The PDF dataset within the EOT 2008 Archive includes examples of each version, from 1.0 to 1.7.

Optimization

PDF documents can be classified by two layouts or file formats: non-linear (not “optimized”) and linear (“optimized”). Non-linear files typically take up less space; however, they are often slower to access because pieces of the document are stored throughout the file. Linear documents store information in a sequential file format and are often referred to as “Web optimized” because they render more quickly in browsers and plugins. The Archive’s PDF dataset included 2,080,602 documents (47%) that were optimized and 2,323,446 documents (53%) that were not optimized for the Web.

Encryption

The PDF format provides the opportunity to encrypt a PDF document in a variety of ways. The PDF dataset (N = 4,404,048) consisted of 4,197,422 documents (95%) that were not encrypted and 206,627 encrypted documents (5%). Encrypted PDF documents impose constraints regarding the actions that users or programs can execute. These constraints include limiting printing, copying, changing, and adding notes. As they age in Web archives, future uses of these files may be limited because of these encryptions.

Embedded Metadata

Creation Dates

Creation dates in PDF files are incorporated in the PDF file itself, which is considered a metafile containing both the objects that comprise the PDF document and information, or metadata, about those objects. Creation dates can either be set by the user or generated by the PDF application. The document creation date for each PDF document was extracted from its metafile. (This date is different from the date that each document was captured.)

One interesting anomaly is that there are a number of examples of “bad data” in the creation date field, such as creation dates set in the future as well as in the distant past. Past creation dates represent the creation date of the intellectual content and not the date of the creation of the PDF. For example, the earliest recorded creation date in the dataset was from year the 1904 with 59 PDFs listing that year as their creation date. The vast majority (93%) of the dataset, (4,113,371 PDF documents) had creation dates between 1995 and 2009, with 2008 being the year most were created.

Surface Area

One metric that emerged as useful for discovering certain classes of content was the surface area of the first page of a PDF document. This was calculated by multiplying the height and width of the document, which were included in the extracted metadata. Items with a surface area of over 500 square inches typically represented maps, posters, and charts. There are 121,490 instances in which the first pages of PDF documents are over 500 square inches in the Archive.

Future Research & Development

This work area demonstrated that is feasible to describe the content of Web archives by format-specific features. Further, it seems feasible to take advantage of these findings to inform the development of search and discovery mechanisms that will aid information professionals in their collection building processes and researchers in their investigations.

Additionally, the methods outlined in this paper could be easily transferred to other file formats, which often include specific characteristics that could be leveraged to provide new views and insights into the content. Examples include indexing of the specific and unique features of image, video, and audio content, which are growing content types in Web archives.

Work Area 4 – Researcher Needs Assessment

The work in this area consisted of a small exploratory study to investigate academicians' research activities so that we could better understand how they might use the contents of the EOT 2008 Archive. Interviews were conducted with researchers in various academic disciplines to (a) determine the range of research areas they study and the research questions they investigate, (b) identify their access needs, and (c) discover new research questions enabled by access to the Archive.

Methods

The particular focus of the semi-structured interviews was researchers' use or anticipated use of the contents in End of Term 2008 Archive. A questionnaire was developed for this purpose as well as to investigate the research process used by academic researchers from research question through data collection and analysis. Additionally, we took this opportunity to elicit researchers' needs and ideas in regard to future content selection for Web archives. The questionnaire was reviewed for content clarity and scope by members of the project team.

Four researchers were selected initially because of their previously expressed interest in the Archive. These researchers identified others whom they thought would be valuable informants. All participants were invited via email to participate in a one-hour interview that was audio recorded. Participants signed a consent form prior to the interview, consistent with the UNT Institutional Review Board approval.

In all, 11 interviews were conducted with researchers at two universities. Table 14 identifies the researchers' academic disciplines grouped under five areas: Humanities (*n* = 3), Social Sciences (*n* = 4), Formal Sciences (*n* = 1), Applied Sciences (*n* = 1), and Natural Sciences (*n* = 2). While three political scientists were interviewed, other disciplines included only a single researcher. With the exception of one Doctoral Candidate, all participants hold doctoral degrees and are university professors.

Discipline	**#**
Humanities	
English	1
History	1
Philosophy	1
Social Sciences	
Economics	1
Political Science	3
Formal Sciences	
Computer Science	1
Applied Sciences	
Journalism - Communication Law	1
Natural Sciences	
Biochemistry	1
Zoology - Aquatic Ecology	1

Table 14. Academic disciplines of participants

The two participants from the natural sciences do not use the Web for their research activities and were not included in this analysis. The content of the remaining nine interviews was analyzed and organized into three main categories: Current Research, Web Archives, and New Research. Current Research includes research areas and research questions currently being investigated, as well as data collection, storage, and analysis. Web Archives includes access needs, data extraction needs, as well as content selection and capture needs. New Research includes research areas, questions, and ideas that are made possible by Web archives. Each of these is discussed in the findings section that follows.

Findings

The findings reveal the wide range of researchers' investigations, needs, and interests. In some areas commonalities across the disciplines emerged, while in others, researchers' needs tended to align by discipline.

Current Research

Research Areas

The participants investigate questions in a range of research areas. Table 15 lists the areas and includes questions that represent their enquiries. A few observations about the research areas are noteworthy:

- Four researchers in three different disciplines research questions related to government policy and law, either national or international.
- Four researchers in three different disciplines research questions that involve the media, and three of these include new media (i.e., Twitter and blogs) in their studies.
- The scope of research for only two researchers, both in the same discipline, is national; the other seven are concerned with questions that are international in scope, either exclusively or in addition to national interest.

Research Area	Representative Research Questions
The Presidency and New Media	What are the differences between traditional news media (e.g., AP) and new media (i.e., online news providers such as foxnews.com and huffingtonpost.com) along several dimensions, including but not limited to headlines, subjects, source, page position, content, tone, and coverage?

Research Area	Representative Research Questions
Political Violence; Foreign Policy; International and Civil Conflict	What is the appropriate role for US foreign policy? How can an intervener in civil conflict shape outcomes? What are the effects of outside involvement in civil conflict? How does the environment impact political protests and political violence? For example, how does water scarcity create competition between communities and the potential for conflict? What is the correlation between government coercion and repression and a number of factors such as: (a) the types of protesters' demands (e.g., ethnic or religious demands versus political demands or economic demands) and (b) the location of protests (e.g., an urban area versus a rural area)?
American Politics: Race, Ethnicity; Institutional, Behavioral, and Elections; Racial and Ethnic Politics (RAP); Woman in Politics	How do minority and female candidates for office differ from other candidates along several dimensions, including: (a) media use (e.g., radio versus television) and (b) the stage in a campaign when money is spent (i.e., in primaries or later in the campaign)? What are the implications of these differences in regard to establishing a level playing field among candidates? How do minorities respond to different media sources? How do public opinions vary across different races and genders?
Science, Technology, and Society: Science Policy and Peer Review	What are the ways in which society both supports and interferes with science? What do science funding agencies, both national and foreign, expect in return for the public funding they grant researchers for scientific research? How do the criteria in peer review systems reflect the expectations of science funding agencies?
19th Century of the American South; Slavery and the Expansion of the United States; Digital Humanities	How did Americans moving west take control of the portion of the continent that Mexico controlled? How did that transfer of power happen? How does the movement of people interact with government structures? How does the flow of available information influence the movement of people?

Research Area	Representative Research Questions
Mathematical Side of Economics: Entrepreneurs; Analytic Model Creation	What factors do entrepreneurs consider prior to starting up their new risky ventures? How can entrepreneurs make this choice better? How can people get out of an undesirable situation and have a decent chance of succeeding as entrepreneurs?
Internet Law – Libel and Copyright Litigation; International Communication - Trans-border Jurisdiction; Media Law and Policy	How do unresolved issues in statutes, such as in the case of multiple personal jurisdictions, affect online journalists or those that publish on web media, such as blogs? How does existing law pertain to new media? How can it be updated? Are certain policies inconsistent with laws? How can policy makers and law makers gel better?
19th Century Literature: Fiction of the Republic of Texas Era; Digital Humanities	Why do fictional texts portray Texas in a substantially uniform pro-Texas manner, despite different authors, different nationalities of authors, and different publishing formats? Why is Texas a topic of fiction on both sides of the Atlantic in the years of the Republic (1836 – 1845)? What is the role of print publications in the Westward expansion of the United States?
Computational Epidemiology; Contagion Model Creation	What are the effects of structure in the population, for example, the hierarchical distribution of ages, on disease outbreak dynamics? Can the presence of disease in different locations be predicted by analyzing existing social media data, such as Twitter content?

Table 15. Participants research areas and questions

Data Collection, Storage, and Analysis

Noteworthy themes about researchers' data collection, data storage, as well as the analytic tools and methods they use emerged. These include:

- All of the researchers use web-published materials in some manner in their research activities.
    - Archived and online newspapers, both current and historical, are primary data sources for researchers in the Humanities and Social Sciences.
    - A few researchers use historical web-published materials available through the Internet Archive to study change over time in entrepreneurship course syllabi and digital humanities research.

    - A few rely heavily on Twitter and blogs for professional communication and access to information and resources related to their research interests.
- Two researchers in different disciplines do not collect data as part of their principal research activity. Rather, they create analytic and predictive models that other researchers can exercise and validate by seeding the models with their own data.
    - Both of these researchers do collect Web-published data for other research interests.
- Two other researchers in different disciplines also do not collect and analyze data, per se. One researcher employs a legal methodology, which is concerned with asserting positions or arguments and evaluating “evidence” (e.g., statutes) not “data”. Another researcher employs a traditional “deep reading” methodology in his research.
- Researchers who do collect data and build databases typically store their data in multiple places. The common storage locations include: Dropbox, external drives, flash drives, cloud servers, and hard drives on computers in separate locations.
- Researchers rely heavily on libraries to digitize resources, create archives and repositories, and provide access services.
- Researchers in the Humanities and Social Sciences need analytic tools, particularly in the areas of optical character recognition, text mining, natural language processing, and topic modeling.
    - A few strive and struggle with varying degrees of success to learn new skills and apply new tools, but most would prefer to rely on computer scientists to build the needed tools.
- The range of analytic tools used by researchers across the disciplines includes:
    - Statistical analysis tools: Excel, SAS, SPSS, STATA
    - Geographic information analysis tools: GIS, ArcGIS
    - Text mining and text analysis tools:
        - Keyword classification tools
        - *Leximancer*: content analysis software
        - *Mallet*: topic modeling software
        - Natural Language Processing (NLP)
        - Optical Character Recognition (OCR) software
        - PDF *‘find’ feature* and WORD *“compare docs” feature:* text change analysis
        - Sentiment analysis tools
        - *WCopyfind* plagiarism detection software
        - Word choice analysis tools

Web Archives

Access Needs

Researchers’ access needs were organized into five categories: (1) Organization and Interaction, (2) Search Capabilities, (3) Viewing Content & Search Results, (4) Information about the Archive, and (5) Information about Content. While there were some common access needs that emerged, a wide range of individual needs were reflected in the findings. A few general observations are noteworthy.

- Researchers are interested in interacting with archived websites along several dimensions that typically include: key policies, issues, events, topics, and government entities.
- Researchers would like search capabilities that combine keyword searches with selected dimensions of the websites (i.e., key policies, issues, events, topics, and government entities).
- Digital humanities researchers are keen to know characteristics of the content (e.g., number of files and file types) so that they can identify the best method of collecting the data and estimate the time and resources involved.

Data Extraction Needs

Data extraction needs roughly align by discipline: Political Science, Digital Humanities, and Computer Science. This alignment is reflected in the following quotes from researchers.

> *"If we can mine 16 terabytes of data with a few lines of code and be able to put that into a spreadsheet format, a tabular format that we can analyze statistically, that's really cool. And some political scientists have the computing skills to do that. Many of us don't. We studied content and how to do the statistics, but not this. It's not our training. It's not what we're trained to do."- Political Scientist*

> *"I need the OCR-text." "There are all kinds of OCR problems." "I need OCR to be improved."*
> *– Digital Humanities Scholars*

> *"What we really need from the Archive is a programming interface that allows our machines to go in there. That needs to be standardized in some way. Have the computation go to the [archived] data, rather than the other way around." – Computer Scientist*

Additionally, a notable point about data extraction in regard to research proposals and funding was made by the economist:

> *"The ability to get the data needed to build the database needed for research in a certain amount of time with a certain amount of effort is a factor in the competitiveness of research proposals."- Economist*

Researchers' data extraction needs were organized in four categories: (1) Pages, Text, Images, and Data, (2) Data about Content, (3) Link Data, and (4) Capture and Change Data. There were some common data extraction needs that emerged, in particular among researchers in the Humanities and Social Sciences. These include:

- Political Science researchers want to have the data they need extracted and given to them in a format they can import into their databases to use with their data analysis tools.
- Researchers in the Humanities and Social Sciences often extract data from PDFs and OCR-text. Extracting the text from images in PDF documents is problematic and the accuracy of text recognition in OCR-text documents is highly variable.

Site Selection and Capture Needs

> *"As a scholarly community we need an ongoing real time capture of the government web presence, in particular for one entity that a researcher could access over time (like whitehouse.gov). A lot of faculty would be interested in identifying the Web content they would like captured." – Political Scientist*

Site selection and capture needs were organized into six categories: (1) Selection of Content to Capture, (2) Frequency of Capture, (3) Depth of Capture, (4) Event-driven Captures, (5) Issue-driven Captures, and (6) Digital Scholarship. A few noteworthy common themes emerged:

- Capturing the content of social media sites, particularly Twitter, and of blogs is of great interest across all the disciplines. Capturing the interactive content, such as comments, is important to researchers in the Digital Humanities and Social Sciences.
- While researchers noted that the inherent international character of social media sites, three researchers in three separate disciplines are interested in capturing non-U.S. websites.
- All of the researchers identified content they would like to have captured; three indicated that having the ability to specify websites for capture was of interest to them.

New Research

> *"There is an entire field in political science called political communication that formally focused on debates, speeches, etc. Political Scientists are trying to figure out how to analyze the Internet, given that it is in flux and there is no established data archive." – Political Scientist*

> *"I think social media, in terms of research interest especially for sociologists and humanists, is going to be hugely important in the next couple of decades. We're going to be really looking back at this moment where [social media] went from something college kids did to where it is foundational for most people." – Digital Humanities Scholar*

This section identifies general research areas and specific research questions that participants thought would be enabled by Web archives, either the EOT 2008 Archive or some other Web archive. Researchers in four disciplines did not specify new areas of research: philosophy, history, computer science, and journalism. However, those researchers did readily discuss future directions in terms of the anticipated impact of web-published data and media, as well as their Web archive access needs. Following are the common research areas identified by researchers in the political science, economics, and English disciplines.

- Research that investigates change over time in several areas:
  - Presidential policy
  - Science policy
- Research that compares various aspects of websites before and after events:
  - Change in presidential administrations: The rhetoric used to describe "what America is"
  - Presidential debates: Changes in whitehouse.gov website content
  - Health events: CDC information published about the threat of SARS as it evolved
- Research that investigates the effects of differential use of languages:
  - Small Business Association publications
  - Safety publications
  - Media use
- Research that compares website content between government entities and departments:
  - White House and Congress: Salient issues included
  - White House and government departments: Foreign policy message
- Research that evaluates the relatedness of content among websites:
  - Centrality of a website
  - Mapping of political communication

Future Research & Development

The findings of the needs analysis strongly indicate that the content in the EOT 2008 Archive is of interest to researchers, particularly in the disciplines of political science and English/digital humanities. Researchers in these disciplines identified a great number of research questions that could be investigated by access to the data in the Archive. The findings also strongly indicate that researchers across the disciplines involved in this research are interested in having the content of the Web captured and archived in support of future research. They are particularly interested in having the content of social media sites and blogs.

It is clear that researchers in most disciplines will need assistance to extract the data they need from the Archive. Researchers will need to identify the content of interest to their research and to specify the data elements and data formats needed in the extracted content. Collaborations between researchers, librarians, information scientists, and computer scientists appear necessary to build the tools that will enable researchers to discover and extract content.

The next step is to take the findings of this exploratory study and validate researchers' needs with a wider group of researchers. In particular working with researchers in political science to develop formal requirements for Web archive access, organization, and mining will be important. Once these requirements are identified, it will be possible for Web archive providers like UNT Libraries and the Internet Archive to develop the tools researchers need.

IV. Project Achievements

1. Papers & Reports[14]
   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. 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.
   c. Web Archive Service Models and Metrics http://research.library.unt.edu/eotcd/wiki/Web_Archive_Service_Models_and_Metric s
   d. Murray, K. & Hartman, C. (2012). Classifying the end-of-term archive. *Archiving 2012 Final Program and Proceedings* (pp. 84-87). Springfield, VA: Society for Imaging Science and Technology. http://research.library.unt.edu/eotcd/w/images/0/0e/murray_classifying_the_endofterm_archive_ist_2012.pdf
   e. Phillips, M. & Murray, K. (2013). Improving Access to Web Archives through Innovative Analysis of PDF Content. *Archiving 2013 Final Program and Proceedings*. Springfield, VA: Society for Imaging Science and Technology.

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. 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
   d. Grotke, A. & Murray, K. (2012, April 2-3). *The United States End of Term Web Archive*. Presented at the CNI Spring 2012 Membership Meeting in Baltimore, MD. Available: http://research.library.unt.edu/eotcd/w/images/8/80/eotproject_CNI_briefing_Spring2012.pdf
   e. Murray, K. & Hartman, C. (2012). *Classifying the end-of-term archive.* Poster presentation at Society for Imaging Science and Technology Archiving 2012 Conference in Copenhagen, Denmark. Available: http://research.library.unt.edu/eotcd/w/images/5/50/eot_poster_archiving2012_krm_24may2012.pdf

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

f. Phillips, M. & Murray, K. (2013). *Improving Access to Web Archives through Innovative Analysis of PDF Content*. Paper accepted for presentation at Society for Imaging Science and Technology Archiving 2013 Conference in Washington, DC.

3. Advisory Board

a. The initial meeting of the board was held at the Library of Congress in December 2009.
b. The second meeting of the advisory board was conducted via conference call in June 2010.
c. The third meeting with the board was held July 23, 2010 in Washington DC. (Note: Gildas Ilien, from the National Library of France (BnF) who was then Chair of the ISO Committee studying metrics for Web Archives (ISO TC46/SC8/WG9) was in attendance.)
d. A final meeting with the board was held November 4, 2011 via Web conference. A presentation reporting the project's findings in Work Areas 1 and 2 was delivered.

4. Subject Matter Experts

a. First meeting: April 25, 2010 in Buffalo, NY. Seven SMEs attended.
b. Second meeting: October 17, 2010 in Washington, DC. All 10 SMEs attended.
c. Third meeting: April 3, 2011 in San Antonio, TX. Twelve SMEs attended, including two new SMEs who had served as arbitrators in the classification exercise.
d. Fourth meeting: October 16, 2011 in Washington, DC. Eleven SMEs attended.