Curation of the End-of-Term Web Archive

Kathleen Murray, Lauren Ko, Mark Phillips; University of North Texas; Denton, Texas, USA

Abstract

The Classification of the End-of-Term Archive research project at the University of North Texas Libraries is investigating the feasibility of machine-generated classification of websites in the 16-terabyte End-of-Term (EOT) Web Archive. The research is being conducted concurrently in two areas: Archive Classification and Web Archive Metrics.

A set of 1,151 URLs within the EOT Archive was analyzed using link analysis methods to identify related groupings or clusters. Investigations into visualization of the underlying relationships among the URLs were also conducted. Subject Matter Experts (SMEs) in the classification of government information manually classified the same set of URLs using the Superintendent of Documents (SuDocs) Classification Numbering System, which is a hierarchical scheme that groups government publications by federal agencies. The SME-classification will serve as the criterion to evaluate the effectiveness of the link analysis.

In a parallel work area of the project, metrics for Web archives were discussed in a focus group with the SMEs, who identified key criteria libraries would likely employ in acquiring materials from Web archives. Participants also identified two service models libraries will need from Web archive service providers: acquisition and access models. A subsequent survey of Federal Depository Libraries measured the demand for each of these models, as well as libraries' perceived capabilities to support long-term preservation and local hosting of materials from Web archives. It appears that some existing library metrics, but more importantly, standard usage statistics will be essential metrics.

Background

As Web archives become available, organizations will seek to include materials from these repositories in their collections. However, such inclusion is often precluded by selection and measurement challenges. The high-level metadata associated with Web archive files does not support material selection in a manner consistent with libraries' collection development policies. Likewise, no standard metrics exist for Web archives, making it difficult to characterize materials in a manner that communicates their scope and value to decision-makers.

The University of North Texas (UNT) Libraries conducted a needs assessment study in 2005 as a part of the Web-at-Risk project, a digital preservation project of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). The study identified major needs and issues confronting librarians, archivists, content providers, and researchers facing the challenges posed by changes in the publication and distribution of government information [1]. Pursuant to findings from that study, UNT Libraries received a research grant from the Institute of Museum and Library Services to address libraries' collection development needs related to government information (Classification of the End-of-Term Archive Project; LG-06-09-0174-09).

Government publications have been organized for over 100 years by the Superintendent of Documents (SuDocs) Classification Numbering System, which is logically extensible to the content in Web archives. Once Web archive content is classified into related groupings, it will become feasible to apply subject analysis to the content and to build information retrieval systems that allow librarians to identify materials for their collections. When these systems are built, archive service providers will need to supply libraries with metrics consistent with existing practices and emerging standards.

To address libraries' needs for selection and metrics, UNT is leveraging its participation in the End-of-Term Web Archive (EOT Archive) project. This important project captured the entirety of the federal government's public Web presence before and after the 2009 change in presidential administrations [2]. The result is a 16-terabyte Web archive of government information.

The research is comprised of two work areas: Archive Classification and Acquisition Metrics. This paper reports the progress the investigation has made to date.

Methods

Archive Classification

Archive Statistics & Sampling

At the outset of the project the EOT Archive was transferred from the Library of Congress to static servers at UNT and the file formats were verified. Subsequently, a Wayback Machine interface to the Archive was created to enable access to known URLs. CDX files were produced using Wayback's warc-indexer and arc-indexer.

Statistics derived from the CDX files revealed that the Archive included 160,211,356 URLs. An early decision was made to limit the classification scope to two domains: .gov and .mil. This resulted in 141,403,247 URLs and 16,016 unique sub-domains. This number of sub-domains remained too large both for: (a) human classification effort on the part of the project's SMEs and (b) effective visualization of the underlying Web graph resulting from the link analysis.

The URLs derived from CDX files were converted to SURT formats and a decision was made to limit the link analysis of the EOT Archive to unique second-level domains, which resulted in 1,647 URLs. Of these, we found that 496 should not be included in the Archive classification, the majority because they would not resolve in the EOT Archive instance of Wayback, which was essential to human classification of the URLs.

The net result was a sample comprised of 1,151 URLs. After extracting the links from the EOT Archive's HTML files, we generated a Web graph of the 62,452 links inter-relating the in-scope URLs. We also designated a weight to each link

representative of the number of actual links (when looking at URLs in their full length form) that existed between each pair of source and target second-level domains. This data was the essential input for subsequent cluster analysis and visualizations of the 1,151 URLs.

Machine Clustering

Using the EOT Archive Web graph for the 1,151 URLs in the sample, four methods were employed to identify clusters. Since the Web graph data reflects the Web sites' inter-relatedness, it was expected that the clusters would identify groups of related URLs. The methods are discussed below.

LinLog Clustering: We used Andreas Noack's LinLogLayout program [3, 4] that positions nodes based on his edge-repulsion LinLog energy model in which each node is assigned a weight based on the sum of the weights of its edges. In our use, we supplied edge weights based on the total number of links between each pair of second-level domains. The program then determines clusters using normalized cut that examines the number of edges linking two disjoint sets of nodes normalized by the total number of edges incident to the nodes in each of the two sets.

Linlog Coordinates With Agglomerative Hierarchical Clustering: We made use of LinLogLayout's force-directed layout technique to map our Web graph to Euclidean space. This produced a pair of x and y coordinates for each node. We then determined clusters using the agglomerative hierarchical clustering algorithm [5] 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.

Normalized Google Distance (NGD): We leveraged the normalized Google distance measure discussed Cilibrasi and Vitanyi [6]. While this is actually a semantic similarity measure, we have found it translates well to our study of link analysis. Typically used to measure the semantic similarity of search terms used in a Google search, the distance between two terms is smaller when the terms are often found on the same page. If the terms are found on separate pages, but never the same pages, then their normalized Google distance is infinite.

In our application of the NGD formula, we measured the distance between government domains based on the similarity of their outlinks.

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \quad (1)$$

Where:

- x and y are domains
- M is the total number of domains in the graph
- f(x) and f(y) are, respectively, the number of outlinks from x and y
- f(x, y) is the number of domains to which both x and y link

Strongest Outlinks and Majority Inlinks. In this method, our starting point is our 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%. Clusters are initialized with a node belonging to a cluster whose centroid is the target to which it is most highly outlinked.

We then determine what to merge by running through centroids of the clusters for multiple iterations until no changes occur. For each cluster, we look at the inlinks of the centroid and the inlinks of the node to which it most highly links (the cluster it would belong to if it wasn't currently named a centroid), which we call the parent. If more inlinks of the centroid name its parent as most highly linked (and are thus in the parent's cluster), the centroid's cluster is merged with its parent's cluster.

For nodes without outlinks weighted above 1%, we have fallen back on initializing them to clusters based on strongest inlink. When there is a tie for most strongly outlinked of a node, the tie is won by whichever competitor shares the greatest percentage of outlinks with the node being assigned.

Visualizations

We experimented with the following visualization tools using different subdomains of the EOT Archive.

- GUESS force-directed visualizations of the National Institute of Health (NIH) subdomain
- Hypergraph visualizations of the NIH subdomain
- Protovis treemap layout visualizations of the Government Printing Office (GPO) subdomain
- Protovis force-directed layout visualizations of eight sub-agencies within the Health and Human Services (HHS) subdomain

Visualizations depicting relationships across the 1,151 second-level domains in the sample included:

- Protovis force-directed visualizations where edge weights were at or above 20% total outlinks for a node, and
- GUESS visualizations of the node positions resulting from running our Web graph data through the LinLogLayout program with all 62,452 edges taken into consideration.

Human Classification

Requirements were specified for a web-based tool to allow the project's ten SMEs to classify the same set of URLs (N = 1,151) as were being investigated by the machine clustering and visualization methods. The classification tool was developed in Django and allowed SMEs to view Websites, assign one or more SuDOCs classification numbers to a site, and add any additional explanatory notes. SMEs could also designate sites as outside the scope of the federal government or within its scope but lacking an author listing within the SuDOCs scheme.

The URLs were randomly assigned to the 10 SMEs and each of 1,151 URL's was classified by two SMEs in accordance with the SuDocs classification system. Three outside arbitrators, with expertise in the SuDocs system, resolved differences in the SME classifications.

Acquisition Metrics

Focus Group Discussion

Early in the project, a focus group discussion was conducted with six of the project's SMEs. Within their academic libraries, these individuals had responsibility for collection development in

the area of government information and had an average of 21 years (*range*: 6-37 years) experience doing so. Five participants were responsible for reporting statistics within their libraries for the materials in their collection(s). Participants reported a range of experience with Web archives, from novice to just above average. None of the participants characterized themselves as experts in this area.

The objective of the discussion was to identify the criteria libraries use in making material acquisition decisions, in particular the countable units that play a critical role in these decisions. The facilitator explained the purpose of the group discussion and encouraged participants to engage in dialog with one another. An audio recording was made of the discussion. At the end of the discussion, participants completed a questionnaire that identified their demographic characteristics, the types of digital content they select for their collections, and their experience with web archives.

The audio recording was transcribed and content analysis of the transcription, as well as the written comments from one SME, identified the key findings. The findings were refined and augmented by the summary notes recorded by staff members who observed the discussion. Descriptive statistics characterized questionnaire responses.

Survey

Findings from the focus group discussion suggested that libraries would primarily be interested in *accessing* materials in Web archives, versus *acquiring* materials for preservation and local hosting. Since the metrics libraries will require from Web archive service providers will likely be different for these two service models, an online survey was conducted to assess libraries' interests in acquiring versus accessing materials in Web archives, as well as to estimate their capability to support acquisition services, such as preservation, hosting, and user access. Additionally, the relationships between three demographic characteristics (depository type, library type, and library size) and libraries' interests and capabilities were measured.

The survey instrument was created, tested, and administered using Zoomerang (http://www.zoomerang.com/), an online survey development and hosting Website. The instrument consisted of nine questions. Prior to initial deployment, members of the project team reviewed the survey instrument for clarity and the project SMEs reviewed it for content validity. The survey instrument was revised based on the review comments.

An invitation to respond to the survey was sent to 1225 libraries in the Federal Depository Library Program (FDLP) and a total of 414 libraries (34%) submitted responses. Included were 22 Regional Depository Libraries and 392 Selective Depository Libraries.

The majority of the participating libraries were academic libraries (n = 316; 76%). Also included were 53 (13%) public libraries, 42 (10%) state and federal government libraries, two special libraries, and one service academy library. The representation by library type is fairly comparable to the representation by library type within the FDLP.

Small, medium, and large libraries were represented, with medium-sized libraries comprising the largest percentage. The participating libraries' sizes were very representative of the libraries in the FDLP.

On average, respondents' indicated they had limited experience with Web archives (*Mdn* = 2; *range*: 1 = novice and 7 = expert). Only four respondents (1%) indicated they were experts, while 151 (38%) rated themselves as novices.

Data was analyzed using IBM SPSS Statistics Version 19. Measures of central tendency were calculated for most questions and, along with descriptive statistics, informed the reclassification of some responses into fewer categories to meet the expected counts in chi square contingency tables. The relationships between the three demographic variables and responses to several survey questions were analyzed using either chi square or Spearman's rho. In all cases, the required significance level was $\alpha = .05$. A standardized residual greater than 2.0 determined contributions to significant results of cells within chi square contingency tables. Lastly, questions 5 and 9 allowed users to submit free-form text responses. These responses were content analyzed to group them thematically.

Findings

Archive Classification

LinLog Clustering

Two sets of clusters resulted from running the LinLog algorithm on the edges where the source and target are both in the EOT Archive. They differ in the manner in which weights on edges were assigned. With the first cluster set, containing 20 clusters, weights on edges were calculated as the ratio of outlinks from a source to a specific target over all outlinks from that source. The second cluster set had 18 clusters and the weights on edges were given as the actual number of occurrences of a link between source and target.

Using the LinLog method, we attained some clusters that are larger than expected. We would have liked to have seen more clusters breaking out from these large groups. We ended up with less than half the number of clusters desired based on the number of top level or parent government author agencies represented in our sample (N = 52).

LinLog Coordinates with Agglomerative Hierarchical Clustering

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 this case the 52 parent author agencies in our sample. The EOT Archive data set reflects the wide diversity of the agencies in terms of size, the number and size of their sub-agencies, and the amount that they publish. Attempts to achieve clusters that are each representative of a single author agency proved quite difficult.

In the initial code to generate clusters, we specified a limit of 60 clusters as the desired result. However, once determined, any clusters made up of a single node were moved to the nearest cluster. This resulted in 59 clusters. After several attempts running this clustering method, we decided not to force single nodes to nearest clusters as there is a good chance they would end up in the wrong cluster.

Additionally, this method depends on the LinLog visualization to produce its clusters. These visualizations, however, are different each time they are produced with the same data set because initial positioning of nodes is random. Two additional cluster sets based on a LinLog visualization were calculated with limits set to 55 clusters and 75 clusters. Each set included the number of clusters specified. This is perhaps our most successful clustering method.

Normalized Google Distance

To experiment, we began with a Web graph of edges between government subdomains that are present in the EOT Archive. Because Google distance looks only at the concurrent presence of two terms and not the number of times they occur together on a single page, we set a threshold of which links to include in our calculations. Thus, we consider only those edges in which the ratio of outlinks from a source to a specific target over all outlinks from that source is greater than or equal to 1%. We assumed that links less common than that are incidental and should not be accepted as an endorsement of one domain by another.

Using this Web graph, we calculated the normalized Google distance between a node and each of the nodes to which it links based on the intersection of their sets of outlinks. For those nodes without outlinks of a ratio of 1% or greater, and for nodes where there is no intersection between their outlinks and those of their outlinked nodes, the cluster grouping could not be assigned with this method, though this occurred for only 3% of the subdomains. When initial distances were calculated, subdomains were clustered with their nearest neighbor (where the Google distance was smallest). Then we continually combined clusters based on their calculated scores until there were no more changes to the cluster groupings. Further work on better utilization of the NGD for clustering needs to be done.

Strongest Outlinks and Majority Inlinks

By initializing with strongest outlinked clusters in this way we eliminated 13 author agencies as centroids. These agencies tend to be relatively small independent agencies with very specific missions. Their dearth of sub-agencies and the meager number of links the agencies themselves have make it difficult to recognize them as independent agencies using this method.

Two author agencies' subdomains were not within the .gov or .mil domain, so they were out of scope from the beginning of our work. An additional 16 second-level domains could not be clustered because no outlink data was available for different reasons, the most common reason being that only the home page was captured in the crawl. This method resulted in 139 clusters, but the clusters look well-related.

Archive Visualization

Typical results for all visualization techniques are available on the Link Analysis pages of the project wiki (http://research.library.unt.edu/eotcd). Many of the graphs are interactive and can be manipulated to enlarge the images or to rearrange the visualizations.

GUESS force-directed visualizations of NIH. We started looking at the inlinks and outlinks of nih.gov because a list of the NIH family of websites was available. Once the graphs were generated, we examined the link relationships of the subdomains within NIH. We learned that looking solely at the number of links between two second-level domains without context is not enough to reliably inform a relationship due, among other factors, to the varying design of websites and the size of organizational divisions. It was decided to consider the ratio of total links to/from a second-level domain to the number of links to/from a specific second-level domain in future visualizations.

GUESS force-directed visualizations of the sample URLs. We also created visualizations of all nodes and links across the set of 1,151 .gov and .mil second-level domains where edge weights were at or above 20% of the total outlinks for a node. (Note: There are 48,000 edges in total.) The algorithm was performed with LinLogLayout.java, then the coordinates were moved over to display in GUESS. Only those URLs in our sample and links that were within our scope (i.e., between nodes in our defined sample) with at least a 1% ratio of outlinks to total outlinks for the source node were displayed. This resulted in 1,133 nodes and 6676 edges being visible. Unlike most force-directed layout algorithms, LinLogLayout takes edge weights into account when laying out the positions of nodes.

Hypergraph visualizations of NIH. The major finding with this method was that it was too difficult to read visualizations that included more second-level domains from the EOT Archive than those representing NIH links. When we did, the resulting images looked like rubber band balls because of the high number of links between nodes.

Protovis treemap layout visualizations of GPO. Interactive treemap layouts relating to GPO subdomains and mimetypes were created. These provided another visualization of the Archive's structure as well as its contents.

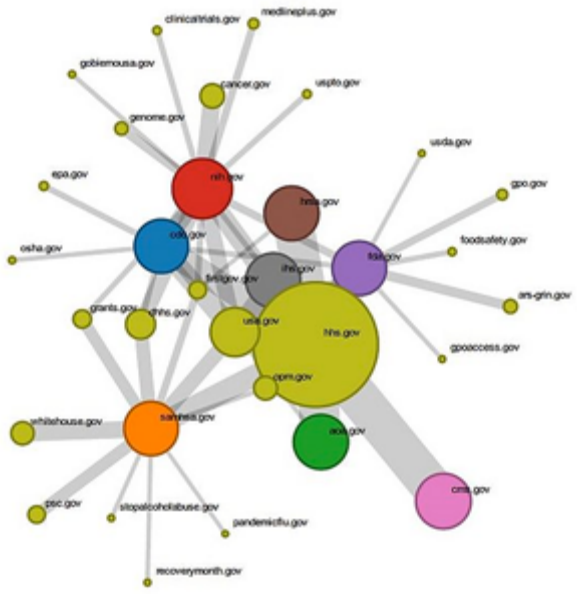

Figure 1. *Protovis Force Directed Visualization of HHS*

Protovis force-directed layout visualizations of HHS. We created visualizations of eight HHS known sub-agencies and the

second-level domains to which over 1% of their outlinks point (Figure 1). The eight sub-agency nodes are colored uniquely. The width of the links is directly related to the percentage of outlinks pointing to the target nodes. The size of the nodes is based on the number of other visible nodes that have links to them, although in cases where edges of two visible nodes have a weight less than the requisite 1%, these edges are not visible.

Human Classification

The project SMEs classified the 1,151 sample URL's in accordance with the SuDocs classification system. Of the 1,151 URLs classified, SMEs were in agreement in 70% of cases (n = 808). For the 30% of URLs (n = 343) about which the SMEs differed in their classifications, the greatest percentage (32%) differed in that one SME included more authors for a URL than another SME, although they were otherwise in general agreement. The next largest disagreements were cases in which one SME assigned an author(s) to a URL and the second found the URL to be within the scope of the federal government but could not find an appropriate author in the SuDoc classification system. In 19% of cases both SMEs classified the URL; however, their classifications were different.

Types and Percentages of Classification Differences

Additional Authors	One of the classifiers added an additional author(s)	32%
Classified v. In Scope but Unable to Classify	One assigned an author(s); second indicated URL was in scope but was unable to classify it	20%
No Agreement	No exact agreement	19%
Parent Author v. Subordinate Office	one assigned a parent author and the second a subordinate office of that parent	16%
Classified v. Out of Scope	One assigned an author(s); second indicated URL was out of scope	10%
In Scope v. Out of Scope	Neither assigned an author; one classified URL as in scope and the other as out of scope	2%

Three arbitrators, with expertise in the SuDocs system, resolved the 343 disagreements. Following the arbitration, all 1,151 URLs in the sample were assigned one or more SuDocs classification numbers. These classifications will inform a set of groupings for the URLs organized by federal agencies. Future work will compare the results of these classification groupings to the results of the link analysis groupings (or clusters).

Metrics

Focus Group Discussion

While the discussion was primarily concerned with electronic resources, there is little doubt that SMEs anticipate materials in Web archives will be akin to electronic resources in terms of the selection and acquisition decisions libraries will make. Web archive service providers will likely need to furnish information that allows libraries to evaluate archived content along the following dimensions:

- Broadness of applicability
- Usage data: vendor-provided & standards-compliant
- Appropriateness for collection
- Number of titles
- Unique content
- Duplicate content

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

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

In addition to their preservation service, Web archive service providers will have the opportunity to provide two additional services for libraries: a hosting/access service and an acquisition service (Figure 2). Findings from the focus group suggest that some libraries will want to acquire materials from an archive, in particular materials that augment the comprehensiveness of a unique collection or materials that are critical to the research focus of academicians. However, access services will be the norm for most libraries, illustrating the need for archives to be positioned to provide standardized usage data.

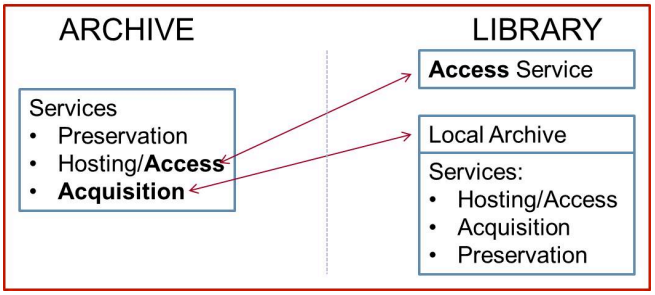

***Figure 2**. Web Archive Services*

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

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

Survey of Federal Depository Libraries

The survey results confirmed what the focus group findings had suggested: Libraries are decidedly more likely to access

materials (*Mdn* = 6; *range*: 1 = extremely unlikely and 7 = extremely likely) than to acquire materials (*Mdn* = 2) from Web archives at trusted institutions (Figure 3). Importantly, libraries have limited capabilities for either long-term preservation of materials acquired from Web archives or for hosting materials for user access (*Mdn* =2; *range*: 1 = not capable and 7 = extremely capable). Libraries' preferences for accessing Web archives is reinforced by their estimates of the support they are likely to receive within their organizations for acquiring materials from Web archives. Just over 60% of libraries of all sizes (*N* = 395) indicated they had either no support (*n* = 86; 22%) or limited support (*n* = 157; 40%) for the acquisition of materials from Web archives.

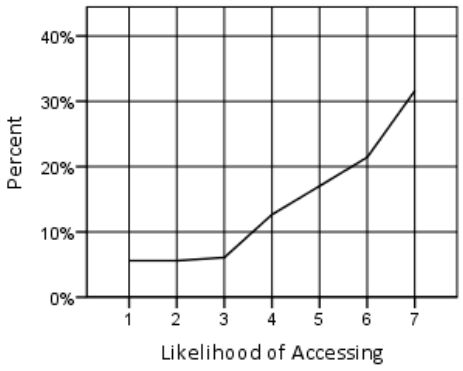

***Figure 3**. Likelihood of Accessing Materials from Web Archives (N = 411; 1 = Extremely unlikely, 7 = Extremely likely)*

Closing

Since the SuDocs classification system is inherently hierarchical by federal agency, it will be straightforward to create groupings or clusters based on the classification numbers assigned by the SMEs. Future work will compare the results of these human-assigned classification groupings to the results from the various machine-generated link analysis clusters. The results from should provide insights that will enable improvements in the accuracy of the machine-generated clusters. It may then be possible to extend the analysis to the remainder of the EOT Archive contents.

Other future avenues of exploration have emerged as a result of the link analysis and visualizations. In the general case, is it possible to characterize certain types of sites (e.g., portals) so that they can be predictably identified within a Web archive? In the specific case of the EOT Archive and the SuDOC classification system: Can we identify additional URLs in the visualizations that are associated with an agency author or a cluster group but are not classified as such? Can we account for URLs that are classified for particular agency authors but do not appear in the visualization as associated with that agency?

Finally, in regard to the metrics for Web archives, while there will be demand from libraries for both acquiring and accessing materials in Web archives, it seems probable that most libraries will have a first preference for accessing materials. With the establishment of standards for usage statistics regarding vendor-provided journal and database services, it may become incumbent upon future Web archive service providers to establish and embrace a complementary set of standards for Web archives.

References

[1] K. Murray and I. Hsieh, "Archiving Web-published materials: A needs assessment of librarians, researchers, and content providers," Government Information Quarterly, 25, 1, (2008).

[2] Library of Congress. "Library partnership preserves end-of-term government Web sites," (2008, August 14). Retrieved March 19, 2011 from http://www.loc.gov/today/pr/2008/08-139.html

[3] A. Noack, "Energy Models for Graph Clustering," Jour. of Graph Algorithms and Applications, 11, 2, (2007).

[4] A. Noack, " Modularity Clustering is Force-Directed Layout, " Phys. Rev. E 79, 026102, (2009).

[5] S. E. Schaeffer, "Graph Clustering," Computer Science Review, 1, 1, (2007).

[6] R. L. Cilibrasi and P. M.B. Vitanyi, "The Google Similarity Distance," (2005). Accessed March 18, 2011 at http://eprints.pascal-network.org/archive/00002784/01/tkde06.pdf.

Author Biography

Kathleen Murray received her PhD in information science from the University of North Texas (2000). She is a postdoctoral researcher in the digital library and Web archiving programs at the University of North Texas (UNT) Libraries. Her work focuses on user studies and collection development.

Lauren Ko received her BA in computer science (2004) and her MS in information systems (2008) from UNT. She is a programmer in the digital library division of UNT Libraries where she conducts research and development in the Web archiving program.

Mark Phillips (MLS, 2004) is Assistant Dean for Digital Libraries at UNT. He directs the UNT Libraries' digital library and Web archiving programs.