Assessment of Need

The audience for this project is any library interested in selecting and acquiring government information from Web archives. As Web archives become more available and accessible, many libraries will be collecting materials from these important information repositories and librarians need the capability to identify and select materials in accord with the mission and scope of their collection policies. At present selecting relevant content from Web archives is a daunting endeavor, in large part because most archives harvest and store web-published materials in a manner optimized for preservation and not in a manner that supports information discovery. Organization of the large amount of content in Web archives in accord with established schemes holds promise as a solution that will enable extension of collection development practices to this new class of materials.

This project specifically addresses collection development needs related to government information, which is included in a wide variety of library collections and for which a well-established classification scheme exists. The Superintendent of Documents (SuDocs) Classification Numbering System has been used to organize government information for over 100 years and is logically extensible to the content in Web archives. It is envisioned that once Web archive content is classified, it will become much more feasible to apply subject analysis to the content and build information retrieval systems that allow librarians to identify and select relevant materials for their collections.

The University of North Texas (UNT) conducted a needs assessment study in 2005 as a part of the Web-at-Risk project[1]. The study identified major needs and issues confronting librarians, archivists, content providers, and researchers facing the formidable challenges posed by changes in the publication and distribution of government information at the international, national, and state levels (Murray & Hsieh, 2008). Pursuant to findings from this needs assessment, this research proposal is concerned with two major needs faced by librarians:

1. Selection of materials in Web archives consistent with collection policies, and
2. Measurement of materials from Web archives for acquisition decisions.

To address these two needs, UNT will leverage its participation in the End-of-Term Web Archive (EOT Archive) project.[2] This important project captured the entirety of the federal government's public Web presence before and after the 2009 change in presidential administrations (Library of Congress, 2008). The result is an approximately 25 terabyte Web archive of government information that is replicated in repositories at the collaborating organizations, including UNT. The EOT Archive will be the information repository used to investigate innovative solutions to the two needs of librarians identified above.

Consistent with these needs, the inclusion of important materials in Web archives, such as the EOT Archive, in collections is often precluded because of selection and measurement challenges. One substantial issue in enabling the selection of government information materials within the EOT Archive is that materials are stored in files containing high-level metadata that does not support selection of materials in accord with the established collection development practices librarians normally use when selecting government information. Likewise, no measurement units or set of metrics for Web archives exists, making it quite difficult for librarians to characterize materials in Web archives in a manner that communicates their scope and value to library administrators and other decision-makers. The proposed research addresses both of these issues.

Collection Development

Federal Government Information. Federal government information is comprised of the information products of the three branches of the United States government disseminated under the auspices of the Federal Depository

[1] The Web-at-Risk project [http://www.digitalpreservation.gov/partners/web-at-risk/web-at-risk.html] funded in 2004 by the Library of Congress under the National Digital Information Infrastructure and Preservation Program.

[2] The EOT Archive is a collaborative project of the Library of Congress, the US Government Printing Office, the Internet Archive, the University of North Texas Libraries, and the California Digital Library.

Library Program (FDLP)[3] as well as information products from other government and non-government sources. Increasingly, government information products are web-born and, in response to this challenge, collection development policies and practices within libraries are changing, particularly in regard to item selection. The Government Printing Office (GPO) is aware of the collection development challenges inherent in electronic collections of government information. Their current policies and practices are not effectively including the spectrum of web-born government information products and GPO offers this guidance to libraries in the FDLP handbook:

> Acquiring electronic information that falls outside the purview of GPO should be factored into any policies or any guidelines you develop as part of an overall collection development policy. Also, the challenges of item selection when developing an electronic collection should be considered when developing your depository library's collection development policies. (FDLP, 2008, p.2)

In established practice, the SuDocs system is used by selective depository libraries to select and organize federal information products for library collections. The EOT Archive consists of web-published federal government information and the SuDocs system is a stable and extensible organizational scheme, widely employed in current collection development practice, which can be used as a structural guide to organize the Archive's contents. Assigning SuDocs classes to the content in the EOT Archive will make it possible to extend existing selection practices employed for government information to the web-published materials in the EOT Archive. Additionally, unlike the live Web, archived materials are preserved in a static state that may lend itself to classification more readily than materials on the live Web.

The Federal Digital System (FDsys). In an effort to better manage both printed and born-digital government information, the Office of Innovation and New Technology of the GPO is overhauling the technical infrastructure for depositing, distributing, and archiving federal government publications that are within the scope of the FDLP. In January 2009, GPO launched the first release of a content management system, FDsys (GPO, 2008).

Harvesting information that is within the scope of the FDLP from government agency Web sites is currently planned for FDsys Release 2, which is not yet scheduled. GPO conducted two web harvesting trials of Environmental Protection Agency (EPA) Web sites in 2006 to identify and capture the rising number of government publications that are published on agency Web sites but are not reported to GPO for inclusion in the FDLP and GPO's Cataloging and Indexing Program (i.e., "fugitive documents"). Significantly, the results concluded:

> In the current state, a great deal of additional manual processing will need to be performed by GPO after content is harvested in order for a given publication to be added to the FDLP collection. . . . Without a technological or financial solution to assist with the additional processing needed, an automated harvesting tool will only move the bottleneck from the discovery and harvest functions into the functions listed above. (GPO, 2007, page 4)

In reaction to these findings, Daniel Cornwall, Head of Information Services for the Alaska State Library, commented: "I was struck by how GPO's identification of intensive processing tasks listed on page 4 mirror our challenges here in Alaska . . . I have to concur with the white paper's assessment." (Cornwall, 2007)

It seems evident that even after the FDSys is fully implemented, collection development challenges will remain for librarians selecting web-published information from either the live Web or from Web archives. GPO recognizes this, and states the importance depository libraries should place on acquiring non-depository materials for their collections.

> The acquisition of non-depository materials becomes more crucial in an electronic environment. With the advent of desktop publishing more and more federal agencies are disseminating information directly to the Internet, thus avoiding GPO altogether. While GPO is engaging in numerous efforts to reconcile this, it

[3] "The FDLP includes all Government information products, regardless of format or medium, which are of public interest or educational value, except for those products which are for strictly administrative or operational purposes, classified for reasons of national security, or the use of which is constrained by privacy considerations." (GPO, 1998, Executive Summary, ¶ 2)

> becomes more incumbent upon you as the depository librarian to exert efforts to capture, retain, and provide access to these electronic materials. (FDLP, 2008, p.10)

It is also important that libraries build collections of government information independently of the FDsys. Government agencies are often subject to the political whim of parties in power to remove information from Web sites, which can extend to Web archives. Libraries are less affected by such influences and have historically protected the integrity of the public record.

Access Issues for Web Archives. Rauber & Masanès (2008) assert that many of the preservation challenges of archiving the Web have found technical solutions and the emphasis of Web archiving research efforts has moved to the challenges inherent in providing access to the archived materials. Two common access applications for Web archives are WERA (Web Archive Access) (Internet Archive, 2007) and the WayBack Machine (Internet Archive, n.d.). Both of these applications support archive searches based on URL and capture date. However, in the absence of a known URL, there is at present no way to access the archived content using these applications.

In combination with NutchWAX, WERA also provides a full-text search engine that works for small to medium Web archives containing about 500 Million documents or around 150k ARC[4] files (Internet Archive, 2007). However, the EOT Archive and many other Web archives exceed these limits and, at present, full-text search functionality fails to meet the tremendous challenges of scale these very large archives pose. Should the scaling issues be surmounted, an indexing/search approach to information discovery and selection would still be inconsistent with established depository library collection development practices (FDLP, 2008) and it seems likely that government information librarians, who now find selection and acquisition of materials in the live Web daunting (Murray & Hsieh, 2008), would encounter similar problems accessing Web archives through a search engine.

The findings of two research projects dealing with browse-related access issues for Web archives were reported at the 2008 International Web Archiving Workshop (IWAW) (Jatowt, Kawai, & Tanaka, 2008; Song & JaJa, 2008). However, these early results do not offer solutions that can be implemented in the near term to address the immediate needs of librarians selecting materials for collections from Web archives.

Classification of materials in Web archives in accordance with an established system, such as SuDocs, has not been researched or trialed. Both search and browse functionality could be better optimized for government information if Web archives were first organized in accord with the SuDocs classification system.

Metrics

Academic Library Statistics. Currently, academic library statistics are gathered biennially by the National Center for Educational Statistics (NCES), and annually by both the Association of College and Research Libraries (ACRL) and the Association of Research Libraries (ARL). The ACRL survey is fairly representative of the three statistical gathering efforts in terms of the traditional statistical reporting areas. In 2007, the ACRL areas included (ACRL, n.d.):

- Collections: which includes monographic volumes, serials and microforms, etc.;
- Expenditures: which includes library materials, electronic serials, computer hardware and software, salaries and wages, etc.;
- Electronic Expenditures: which includes computer files; electronic serials; internal and external bibliographic utilities, networks, and consortia; computer hardware and software; and document delivery / ILL;
- Personnel and Public Services: which includes size of staff, reference transactions, circulations, ILL, etc.;

[4] ARC: “Specifies a method for combining multiple digital resources into an aggregate archival file together with related information, used since 1996 by the Internet Archive to store 'web crawls' as sequences of content blocks harvested from the World Wide Web.” Retrieved January 15, 2009 from http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml

- Ph.D.s granted, Faculty and Enrollment Statistics: which includes Ph.D.s granted, number of faculty, undergraduate and graduate enrollment, etc.;
- Supplementary Statistics For Networked Electronic Resources and Library Digitization Activities

Unlike the NCES survey, the ACRL does collect supplementary statistics for digitization activities: Digital Collections (number of collections, size, and items); Usage (number of items accessed and number of queries); Direct Costs (personnel and equipment/software/contract services); and Volumes Held Collectively. However, with the exception of these supplementary statistics, the statistical collection areas and specific measurements do not translate well, or often at all, to materials harvested to a library's Web archive.

Within its Statistics and Measurement Program, the ARL conducts three surveys of its 123 member libraries (ARL, 2009). One survey collects statistics in areas that roughly correspond to the ACRL and NCES survey areas and a second survey is concerned with preservation. The ARL also gathers detailed data from its member libraries regarding digital library holdings and services in its Supplemental Statistics survey. Additionally, the ARL has developed a new measurement instrument, DigiQual, to assess the quality of services for digital library users. The ARL has clearly gone beyond other measurement efforts in the area of digital libraries; however, they do not provide either equivalencies or new statistical measures for Web archives.

Metrics for Web Archives. Librarians are often at a loss to communicate the scope and value of Web archives to the administrators within their libraries in terms these decision makers can readily understand (Murray & Hsieh, 2008). It would be helpful to have equivalencies for Web archives that parallel physical and digital library statistics and measures. Martell (2008, p. 405) identifies a similar issue in his discussion of the decline in the physical use of academic libraries and the increased usage of electronic resources:

> If only we had equivalences for physical use (e.g., using a print article on site) relative to electronic use (e.g., using an electronic article off site), then we would know the cost implications for each type of use. It would be fascinating to know how many of one it takes to equal one of the other.

While measures for digital libraries and services are emerging, measures for Web archives are woefully lacking, as is evidenced in a recent survey of the members of the International Internet Preservation Consortium (IIPC) (Grotke, 2008). The survey asked: "What statistics do you use when reporting to management or others in your organization about web archiving activities?" Of the 33 responses, the two most common statistics were size (terabytes, gigabytes, etc.) (63.6%, *n*=21) and number of Web sites or other Internet resources selected (54.5%, *n*=18). In a group whose members are rightly considered leaders and pioneers in Web archiving, it appears that most have yet to go beyond two basic measurements of their Web archives: "how much" and "how many".

In the *Web Archiving Metrics* session of the 2008 IIPC General Assembly Meeting, Kunze (2008) suggested three reasons for developing a set of common metrics for Web archives: to characterize the analysis of Web archives, to enable planning for Web archives, and to justify resources for Web archives. He suggested there was a need for "human understandable archival units" of measure versus the more commonly reported measures of size (terabytes) and counts (URLs or hostnames), which often fail to characterize Web archives in a manner that is understandable to organizational management. Analogously to Martell's desire for "equivalences" for physical and electronic usage, Kunze wondered if converting measures of archived Web content to physical units might be a fruitful strategy to develop credible archival units that provide a basis for comparison among archival efforts.

The concept of equivalencies between units of measure for Web archives and commonly understood physical units holds promise for meeting the needs of librarians to both characterize and quantify materials in Web archives to library administrators and decision-makers. The EOT Archive is a test platform for identifying a set of metrics for Web archives.

Project Goals and Benefits

To assist librarians in meeting the collection development challenges posed by Web archives, the goals of this project are: (a) to effectively classify the EOT Archive in accord with the SuDocs classification scheme and (b) to identify a set of metrics for Web archives that effectively communicate their scope and value. Achieving these goals will address the librarians' needs in regard to selection and acquisition of materials in Web archives.

In a unique approach to the organization of Web archives, this project proposes to classify the materials in the EOT Archive in a manner that will set the stage for future information retrieval systems to extend current collection development practices to archived Web sites. This proposed classification will enable government information librarians to select SuDocs-classified materials that are critical to their mission and reflected in their current collection policies as well as to more readily identify new materials that have not been distributed through the FDLP. As depository libraries are better able to fulfill their traditional mission of assuring public access to all government information, including web-published information, the public will be the ultimate beneficiary.

Additionally, in response to the deficit of established metrics for materials in Web archives, this project proposes to identify metrics that fill the gap. The result will enable librarians to apply metrics that translate the measurable units appropriate to materials in Web archives to units more familiar to libraries and their administrations. This need is evident not only for government information librarians but for academic libraries and potentially for any library seeking to add materials from Web archives to their collections.

National Impact and Intended Results

Web archiving research has moved beyond an emphasis on preservation issues, to include issues related to accessing the materials in Web archives (Rauber & Masanès, 2008) and to creating common metrics for Web archives (Kunze, 2008). This research project is unique in investigating a classification approach to the access challenges. Likewise, only preliminary investigations regarding the creation of metrics for Web archives have been undertaken. Both material selection and measurement are integral to collection development practice in libraries. Thus the results of this research directly transfer to the larger library community seeking to include Web archives and their contents in collections.

Access to materials in Web archives has generally been provided by limited search and browse functionality that takes advantage of the high-level, machine-generated metadata available for captured Web sites.[5] This research proposes a classification approach to organizing Web archives that has never been undertaken and yet is well-suited to government information collection development practice. It is expected that the classified EOT Archive will facilitate the extension of current selection practices used by government information librarians to materials in Web archives. The resulting organizational structure will also lay the necessary groundwork to enable development of information retrieval systems that optimize material selection in accord with existing classification systems and greatly improve the relevancy of their search results. Librarians in virtually any subject area will benefit from government information librarians being able to more readily identify materials of interest from government information archives, such as the EOT Archive.

If this research is successful, the stage will be set to test whether this approach can be generalized to other government information Web archives at the state and international levels, which generally classify government information using a well-defined system. (See Appendix A.) Further, non-government classification systems based on organizational or publication structures (versus intellectual content) could be used for organizing Web archives, thereby supporting collection development practice across multiple domains. It is significant to note that members of the IIPC have been harvesting the Web space of their national governments in whole or in part for several years. Through its membership in the IIPC, UNT Libraries is well-positioned to leverage the results of this classification approach to the international community.

This research anticipates the day when libraries will be selecting and acquiring materials from Web archives and will need to characterize these materials using metrics that are commonly understood and reported by academic libraries. At present, many libraries grapple with describing and quantifying materials in Web archives in a manner that both substantiates their significance for a library's collection and quantifies the resources in terms familiar to library administrators and funding organizations. Considering the amount of government information being published on the Web and outside the scope of the FDLP, this is a particularly significant problem for government

[5] This is not true for human-curated Web archives, which are quite resource-intensive. With the proposed classification approach, it seems likely that resource-intensive access issues identified by the GPO in their EPA Web Harvesting Whitepaper (GPO, 2007) can be avoided by providing access to the web-published materials in their native or web-born form, that is, without additional effort to extract, repackage, and apply structured metadata.

information librarians. The current practice of statistical measurement for academic libraries offers no solutions. In the wider Web archiving community, statistics for Web archives are inconsistently collected by libraries and lack common metrics (Grotke, 2008; Kunze, 2008). Establishing meaningful metrics for Web archives is a critical step in establishing the value of Web archives and facilitating their use and acquisition by libraries. Metrics will shed light on the value of Web archives and promote their use by collection development professionals.

This research proposes to identify a set of metrics for Web archives that will be informed by government information librarians and tested for materials they select from the EOT Archive. The results will provide a method for translating the little-understood phenomena of Web archives into familiar terms for librarians and administrators. This should promote the inclusion of web-harvested materials in core library collections, a practice whose importance will increase as more materials are web-born and web-published. Researchers and managers of Web archiving projects are eager for common metrics to communicate the scope and value of their Web archives, both to decision makers and funding agencies in national governments and at leading research universities. UNT Libraries is well-positioned to leverage the results of this research on metrics for Web archives to the national and international Web archiving community.

The Digital Project Unit (DPU) at the UNT Libraries builds and maintains its Web archives and digital collections, including the EOT Archive, on open-source platforms. The processes and tools developed to address this project's research goals are expected to be applicable in other open source Web archive environments. Likewise, the Web archiving tool set used within the UNT Libraries conforms to standard, open source practices within the national and international Web archiving community. Thus, other Web archives should be able to adapt and apply the methods and tools developed in this project. All processes and tools created in this project will be documented and publicly available. Additionally, project results will be shared in appropriate Web archiving forums with other research libraries and organizations. There is no doubt that the results of this study will be of interest to many academic, state, and national libraries and to the international Web archiving community. Finally, the results of this research will provide a basis for further web archive organizational and access studies, as well as offering a set of metrics for a new class of materials whose importance for collection development will continue to grow.

Project Design and Evaluation Plan

Research Questions

1. How effective is the organization of large-scale unstructured Web archives using a pre-defined classification system, the SuDocs classification numbering system, as evaluated by government information librarians?
2. What measurable units for the materials in Web archives best support management acquisition decisions in libraries?

Plan of Work

Participants – Advisory Board

Participants in this study will be 10 government information librarians from depository libraries around the United States. They will be recruited by the principal investigator, who is a librarian with several years of professional experience as a government information librarian at a selective depository library and who served as chair of the Federal Depository Library Council[6]. Participants will serve as Subject Matter Experts (SMEs) in the area of collection development for government information. They will also serve as Advisory Board members actively involved in the project.

[6] The Depository Library Council consists of fifteen members appointed to three-year terms by the Public Printer. The Council's mission is to "assist the Government Printing Office in identifying and evaluating alternatives for improving public access to government information through the Depository Library Program (DLP) and for optimizing resources available for operating the Program" (GPO, 2005).

Scope of Work

The project is comprised of two work areas: Archive Classification and Acquisition Metrics (Figure 1). Work Area 1 will employ Web structural analysis methods to identify Web site groupings within the EOT Archive and measure the effectiveness of this method in classifying the EOT Archive's content by comparing the resultant groupings to a classification map created by the study participants. The map will match the EOT Archive's URL seed list[7] to the SuDocs classification numbering system (SuDocs, 2008). Work Area 2 will employ both qualitative and quantitative analysis to identify key measureable indicators of the value of materials contained in Web archives. It is anticipated that these metrics will support librarians' acquisition decisions.

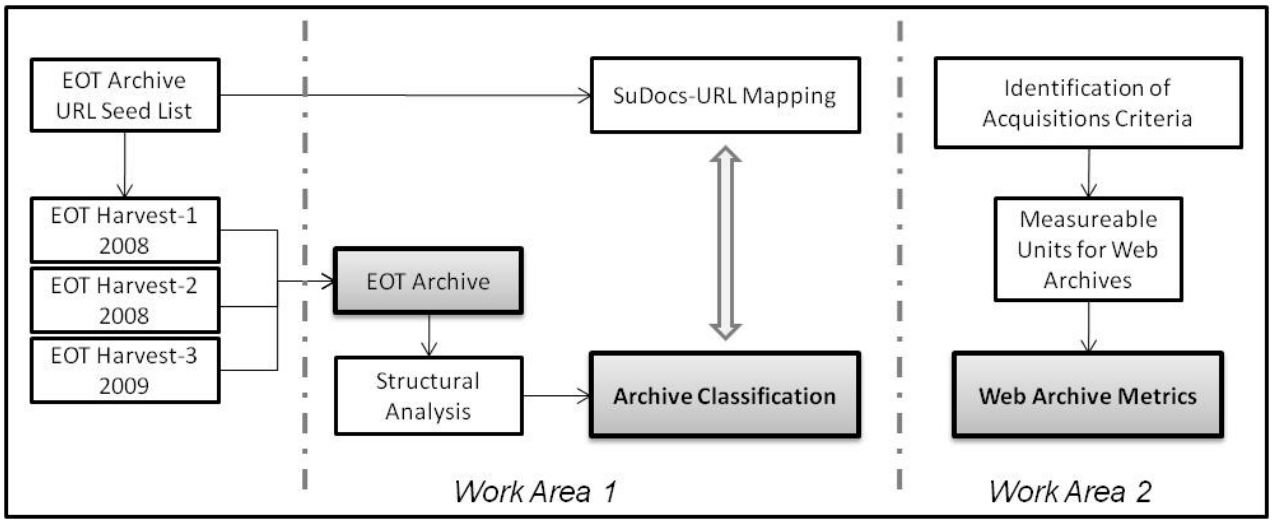

Figure 1. Project Work Areas

Work Area 1. EOT Archive Classification

1.1 Data Storage Environment for Research and Testing

For the completion of this grant, a significant amount of storage will be required for the research and testing environment. The UNT Libraries is storing a preservation copy of the EOT Archive in its archival repository, but a duplicate copy must be created for the research and testing this project proposes. The hardware expenditure is for a 48 TB storage server and two access nodes. The server will store a copy of the EOT Archive, the tools and applications created during the project, and the collections identified by the SMEs in 2.3. The access nodes will add needed computing resources and will provide access to the Archive for researchers and librarians.

1.2 Structural Analysis

The Federal government's Web presence for the EOT Archive was defined by a list of Web site URLs, which were contributed by the Internet Archive, the California Digital Library (CDL), and a group of SMEs using a URL nomination tool developed by UNT. Within a Web archive, web-harvested materials are stored in files (ARC/WARC formatted files) containing high-level metadata that does not support identification and selection of materials in accord with a classification system, such as SuDocs. The captured URLs within the EOT Archive will be initially analyzed by programmers on the UNT project team using a link analysis tool developed by the Internet Archive, whose staff will provide consultation as needed for successful application of the tool. The analysis will be augmented through enhancements to the tool developed by programmers on the UNT project team, again in consultation with the Internet Archive staff. The link analysis will yield Web graphs of the linked relationships among the seed URLs, both links out and links in. These relationships should reflect topical relationships among the

[7] A *seed list* includes one or more URLs from which a web crawler begins capturing web-published materials. Web crawlers extract additional candidate URLs for capture from the web pages in the seed list (Hsieh, Murray, & Hartman, 2007).

archived Web sites as well as indicate the relative importance of the Web sites. Visualization of these relationships will depict a topical organizational structure of the EOT Archive.

As needed for the identification of major Web site groupings or sub-networks, measures from social network analysis, for example k-cores, will further analyze the graphical data from the link analysis (Ortega and Aguillo, 2008).Likewise, co-link analysis will augment and further depict the relationships among the seed URLs (Ortega and Aguillo, 2008; Zuccala & Thelwall, 2007) to aid in identification of the Web site groupings within the Archive.

1. 3. Mapping the EOT Seed URL List to the SuDocs Classification Numbering System

The SuDocs system is a hierarchical system, in which subordinate entities are classified under top-level government entities (*N*=59). For example, the Department of Agriculture is a top-level entity that includes several subordinate entities, such as the Forest Service. To establish a standard that will serve as a criterion measure for the structural analysis classification of the EOT Archive, participants will map the URLs (*N*=2,400) in the EOT Archive's seed list to the author level of the SuDocs classification numbering system (*N*=560). The author level identifies government authors for "each executive department and agency, the Judiciary, Congress, and other major independent establishments" (SuDocs, 2004).

Each URL will be classified by two participants in order to measure inter-rater reliability of the classification. Project programmers will create a web-based classification application to facilitate mapping the URLs. This application will be created using the Django Web framework which has been deployed successfully in many other production applications created by UNT. In order to optimize the reliability of their classification, a training session for participants will be held to (a) instruct them in using the web-based classification application, (b) provide necessary background information regarding Web archives, and (c) provide a classification practice session to provide any clarification. The resulting classification map, the *SME Map*, will serve as the standard against which the effectiveness of the structural analysis method of classification described in 1. 2 will be measured.

1. 4. Classification of Web Site Groupings

Each grouping resulting from the structural analysis will contain a list of URLs. A representative sample of the Archive's groupings will be selected for manual classification.[8] The project team at UNT will inspect the actual Web sites within the sample and classify their government agency authors in accord with the SuDocs classification numbering system. This will result in a mapping, the *Structural Map*, of the URLs (including seed URLs) in the sample to the SuDocs system. The expectation is that the discrete groupings, or sub-networks, of Web sites within the EOT Archive, as identified from the structural analysis will correspond to the government classes and, hopefully, to the government authors listed in the SuDocs classification numbering system. Further, it should be possible to identify Web sites that are related to government authors but are not listed in the SuDocs system

1. 5. Evaluation Plan for Work Area 1

The classification maps developed by the participants in 1.3 will be compared to the classification maps resulting from the structural analysis in 1.2 using a side-by-side analysis method. Effectiveness of the structural analysis method for SuDocs classification of the Web sites in the EOT Archive will be measured by the percentage of agreement between the SuDocs classification maps for the seed URLs and the structural classification of the subset of groupings. Discrepancies between the two classification maps will also be identified. Additionally, Web sites not corresponding to SuDocs government authors will be identified. Finally, any problems encountered and limitations of the analysis will be identified.

Work Area 2. Web Archive Metrics

2.1. Identification of Acquisitions Criteria

A focus group will be conducted with project participants to identify the criteria their libraries use in making material acquisition decisions, in particular the countable units that play a critical role in these decisions. Key attributes of

[8] Automatic classification is a logical next step if the research results indicate the overall classification approach used in this study is effective.

each criterion will also be identified. The group discussion will be informed by a pre-analysis of criteria from existing guidelines and practices, such as those published by statistics and measurements programs with the Association of College and Research Libraries (ACRL) and the Association of Research Libraries (ARL).

2.2. Determination of Web Archive Measurement Units

The criteria identified in Step 1 of this work area will be operationally defined by measurable characteristics of the materials in the EOT Archive. Measurable characteristics might include Web site attributes ranging from the number of files by types to aggregate file sizes for a set of URLs. The result will be a metric tool based on a set of rules that translate Web archive materials into measurable units that will assist libraries in making management acquisition decisions for these material types.

2.3. Acquisitions Exercise

In order to test the metric tool developed in 2.2, an acquisitions exercise for each participant's library will be done. UNT will prototype an extraction tool to select materials consistent with participants' current collection development policies, specifically their SuDocs selection lists, from the SuDoc classified EOT Archive. The selected materials will form experimental collections of archived materials for which measurable units will be determined. Extracting measurement data from the ARC/WARC files for the many material formats in the EOT Archive will require custom programming to derive the appropriate metadata. UNT expects to use available software tools when possible, but there will be a significant amount of programming completed during the project. Programmers will investigate the use of parallel programming paradigms such as Hadoop to address issues that may arise with data extraction and with indexing on such a large scale. Resulting code will be released as open source tools whenever possible.

2.4. Evaluation Plan for Work Area 2

Participants will evaluate the respective acquisition metrics for their collections and through a focus group discussion identify ways in which the acquisition metrics could be improved. Subsequently, a report will document this process and the resulting measurement units.

Project Resources: Budget, Personnel, and Management

Budget. The attached budget represents the necessary funding to accomplish our proposed research. The requested funds from IMLS cover the salary of Dr. Kathleen Murray who will design, administer, and analyze the research gathered for the project. In addition, a software programmer will be hired to work on the classification tool for the SMEs and to run processes on the structural analysis of the dataset. Additional expenses and matching funds are discussed fully in the budget justification.

Personnel. The principal investigator at UNT is Cathy Nelson Hartman, Asst. Dean, Digital and Information Technologies. Hartman has extensive experience successfully managing grant-funded projects. Dr. Kathleen Murray served as the Assessment Analyst for Web-at-Risk project,[9] and currently conducts research for the IMLS funded project, *Optimizing the User Experience in a Rapid Development Framework.*[10] Mark Phillips, Head of the Digital Projects Unit, holds comprehensive experience in managing open-source software development projects, as well as optimizing work-flows within the digital lab. Phillips participated in the Experimental Path of the Web-at-Risk project and manages open source software development for the *Optimizing the User Experience in a Rapid Development Framework* research project. The software developer to be hired for this project will develop the classification tool and run structural analysis processes on the EOT Archive.

UNT Libraries was a pioneer in building the CyberCemetery, a collection of websites of government agencies and commissions selected because they ceased operation and to prevent permanent loss of their information content. Building on this experience, the Libraries partnered with the California Digital Library and New York University in

[9] A collaborative partnership of the California Digital Library, the University of North Texas, and New York University, funded by the Library of Congress as part of the National Digital Information Infrastructure and Preservation Program (NDIIPP). (http://Web3.unt.edu/Webatrisk/delivs.php)

[10] IOGENE: "Interface Optimization for Genealogists" (http://iogene.library.unt.edu/)

the Web-at-Risk project to create a web archiving service to allow librarians and archivists to build and manage selective collections of web-published materials. As part of the Web-at-Risk project, UNT Libraries created guidelines for developing collections of web-published materials (Hsieh, Murray, & Hartman, 2007).

Management. The University of North Texas possesses sufficient resources, financial management skills, and technical expertise to successfully implement this project and achieve these research goals. Cathy Nelson Hartman will provide overall management of the project and handle all contracts and budgets. Dr. Kathleen Murray will conduct the research for this project by designing, administering, and evaluating all of the assessment work. Dr. Murray will report and measure the findings and outcomes from the research gathered in this project as outlined in the schedule of completion. Cathy Hartman will coordinate the two annual meetings with the SMEs, as well as communicate with them regularly via conference calls and the project website. Mark Phillips will oversee all technical development and processes. The UNT Libraries project team will meet regularly to ensure completion of the project goals and objectives.

Partner Organizations

Internet Archive. For this project, the Internet Archive will provide open source tools and technical support to UNT. This includes web link analysis and other tools developed by Internet Archive to mine, analyze, and graph web data. The Internet Archive will provide consultation on the use and extension of the tools, as well as feedback on the key findings of the project. UNT will provide programming and technical resources to test and enhance the tools. UNT will have final responsibility for development and application of the tools. Resulting enhancements will be available to both partners as well as being made publically available.

Dissemination

Project information will be disseminated in various forums to reach intended audiences, including a project Web site and blog to chronicle the research and development process. Tools and methods resulting from this project will be publically available and will be shared with IIPC members. Presentations will be submitted to Web archiving, digital preservation, and digital library conferences, and professional meetings, such as the International Web Archiving Workshop (IWAW), the International Conference on Preservation of Digital Objects (iPRES), and the Joint Conference on Digital Libraries (JCDL). Articles will be submitted to appropriate professional publications, such as *Government Information Quarterly* and *Documents to the People.*

Sustainability

The UNT Libraries is committed to the development of digital collections and Web archives as a core function of their services, research, and outreach. Distinguished for its exemplary leadership in creating the CyberCemetery and other government-related digital collections, UNT exists as one of only nine Affiliated Archives of the National Archives and Records Administration (NARA). The UNT Libraries is committed to preservation of the EOT Archive and will continue exploring research questions and demonstrating access solutions to the EOT Archive. The UNT Libraries also participate in the Texas Digital Library, a digital infrastructure and institutional repository to support the scholarly activities of Texas universities. The UNT Libraries serve as a key partner in the Texas Heritage Digitization Initiative providing integral services, hosting of assets and software development. Participation in the NDIIPP funded Web-at-Risk project illustrates UNT's commitment to following best practices for the preservation and sustainability of digital assets.

In 2003, the Libraries created the Digital Projects Unit (DPU) to support these programs, as well as other digital initiatives within the Libraries. The DPU's staff includes six librarians, three programmers, three administrative positions, and a variety of skilled graduate and student workers. The DPU receives further support from its role within the Information Technology Services division at the Libraries, a division staffed by 30 full-time employees. Grant funding from a variety of sources is also received. (See Appendix B).

References

Association of College and Research Libraries [ACRL]. (n.d.) *Academic library trends and statistics.* Retrieved January 16, 2009 from http://acrl.telusys.net/trendstat/2007/index.html

Assocation of Research Libraries [ARL]. (2009). *Statistics and measurements*. Retrieved January 16, 2009 from http://www.arl.org/stats/

Cornwall, D. (2007, February 23). FDSys [Comment 1]. Comment posted to https://www.blogger.com/comment.g?blogID=10013556&postID=117155514090502468

Federal Depository Library Program [FDLP]. (2008, November 18). *Chapter 5: Depository collections.* Retrieved January 16, 2009 from http://www.fdlp.gov/component/content/article/119

Grotke, A. (2008, December 16). *The International Internet Preservation Consortium 2008 member profile survey results.* Retrieved January 22, 2009 from http://netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf

Hsieh, I. K., Murray, K. R., & Hartman, C. N. (2007). Developing collections of Web-published materials. *Journal of Web Librarianship, 1*(2), 5-26.

Internet Archive. (2007, November 6). *Nutchwax: Frequently asked questions.* Retrieved January 14, 2009 from http://archive-access.sourceforge.net/projects/nutch/faq.html

Internet Archive. (n.d.) *WayBack machine.* Retrieved January 14, 2009 from http://www.archive.org/web/web.php

Jatowt, A., Kawai, Y., & Tanaka, K. (2008, September 18 & 19). *Using page histories for improving browsing the Web.* Paper presented at the 8th International Workshop on Web Archiving - IWAW 2008, Aarhus, Denmark. Retrieved January 12, 2008 from http://iwaw.net/08/IWAW2008-Jatowt.pdf

Kunze, J. (2008, April 7). *Web archiving metrics session.* Unpublished presentation, IIPC General Assembly Meeting, National Library of Australia, Canberra, Australia.

Library of Congress. (2008, August 14). *Library partnership preserves end-of-term government Web sites.* Retrieved January 9, 2008 from http://www.loc.gov/today/pr/2008/08-139.html

Martell, C. (2008). The absent user: Physical use of academic library collections and services continues to decline 1995-2006. *Journal of Academic Librarianship, 34*(5), 400-407.

Murray, K. R., & Hsieh, I. K. (2008). Archiving Web-published materials: A needs assessment of librarians, researchers, and content providers. *Government Information Quarterly, 25*(1), 66-89.

Ortega, J. L. & Aguillo, I. F. (2007). Visualization of the Nordic academic web: Link analysis using social network tools. *Information Processing & Management, 44,* 1624-1633.

Rauber, A. & Masanès, J. (2008, November/December). Report on the 8th International Workshop on Web Archiving - IWAW 2008. *D-Lib Magazine, 14*(11/12). Retrieved January 12, 2009 from http://www.dlib.org/dlib/november08/rauber/11rauber.html

Song, S. & JaJa, J. (2008, September 18 & 19). *Fast browsing of archived Web contents.* Paper presented at the 8th International Workshop on Web Archiving - IWAW 2008, Aarhus, Denmark. Retrieved January 12, 2008 from http://iwaw.net/08/IWAW2008-Song.pdf

Superintendent of Documents, U.S. Government Printing Office [SuDocs]. (2004, May 24). *An explanation of the Superintendent of Documents Classification System.* Retrieved December 23, 2008 from http://www.access.gpo.gov/su_docs/fdlp/pubs/explain.html

Superintendent of Documents [SuDocs]. (2008, July). *List of classes of United States government publications available for selection by depository libraries.* Retrieved December 15, 2008 via GPO Access: http://www.access.gpo.gov/su_docs/fdlp/pubs/loc/2008july.pdf

United States Government Printing Office [GPO]. (1998, October 1). *Managing the FDLP electronic collection: A policy and planning document.* Retrieved January 24, 2009 from http://www.gpo.gov/su_docs/fdlp/pubs/ecplan.html

United States Government Printing Office [GPO]. (2005, October 5). *Depository library council: About.* Retrieved January 16, 2009 from http://www.gpo.gov/su_docs/fdlp/council/aboutdlc.html

United States Government Printing Office [GPO]. (2007, February 14). *Web harvesting white paper.* Retrieved January 9, 2009 from http://www.fdlp.gov/repository/web-harvesting/web-harvesting-white-paper/download.html

United States Government Printing Office [GPO]. (2008, April 25). *GPO's Federal Digital System (FDSys).* Retrieved January 14, 2008 from http://www.gpo.gov/projects/fdsys.htm

Zuccala, A. & Thelwall, M. (2007). Web intelligence analyses of digital libraries. *Journal of Documentation, 63*(4), 558-589.

Referenced Web Sites

- Arizona State Library: The Persistent Digital Archives and Library System, or PeDALS
  http://www.pedalspreservation.org/
- California Digital Library: Web-at-Risk Project
  https://wiki.cdlib.org/WebAtRisk/tiki-index.php
- CyberCemetery
  http://govinfo.library.unt.edu/default.htm
- Federal Digital System (FDsys)
  http://fdsys.gpo.gov/fdsys/search/home.action
- International Internet Preservation Consortium
  http://netpreserve.org/about/index.php
- Internet Archive
  http://www.archive.org/index.php
- National Digital Information Infrastructure Preservation Program (NDIIPP)
  http://www.digitalpreservation.gov/
- University of North Texas Libraries' Digital Projects Unit
  http://www.library.unt.edu/digitalprojects/