Improving Access to Web Archives through Innovative Analysis of PDF Content

Mark Phillips and Kathleen Murray; University of North Texas Libraries; Denton, Texas, USA

Abstract

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 five partners identified key resources from the U.S. .gov Top Level Domain and completed crawls from September 2008 until March 2009. The resulting End of Term (EOT) 2008 Web Archive, a 16 TB dataset, was distributed to partners interested in providing local services and access to the archive. The UNT Libraries investigated Portable Document Format (PDF) files, 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. The paper discusses the overall workflow and describes the tools used to extract document features. Findings suggest opportunities for the development of retrieval tools that will provide new ways of selecting content and building collections from large Web archives.

Background

As Web archives become more available, organizations will seek to include materials from these repositories in their collections. However, such inclusion is often precluded by content identification and selection challenges. This is in part because the high-level metadata associated with Web archive files does not support material selection in a manner consistent with libraries' collection development policies. To address this problem, the University of North Texas (UNT) Libraries conducted a needs assessment 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) [1]. The study identified collection development needs and issues confronting librarians, archivists, content providers, and researchers who deal with the challenges posed by changes in the publication and distribution of U.S. government information. A number of government information professionals identified the PDF format as being of significance in their collection development processes. In fact, for many professionals PDF-formatted documents were the unit they were most interested in capturing during the Web archiving process [2].

In 2009, UNT Libraries received a research grant from the Institute of Museum and Library Services (IMLS) to continue investigating libraries' collection development needs relative to Web-published government information (Classification of the End of Term Archive Project; IMLS LG-06-09-0174-09) [3]. UNT leveraged its participation in the End of Term Web Archive (EOT 2008 Archive) project, a collaborative effort of the Library of Congress, the Internet Archive, the California Digital Library, the U.S. Government Printing Office, and the University of North Texas [4]. This important project captured the entirety of the federal government's public Web presence before and after the 2009 change in U.S. presidential administrations. The result is the 16-terabyte EOT 2008 Archive containing 160,211,356 URLs [5]. The largest Top Level Domains (TLDs) are listed in Table 1 and the top four file formats by number of mime-type are listed in Table 2.

Table 1. Number of URLs & Subdomains by Top Level Domains

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

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 overarching question directing this investigation 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.

Table 2. EOT Archive Mime-types by Number of Files

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

Methods

Essentially three steps were involved in this investigation. The first step employed a set of processes to extract a dataset of complete, or fully-renderable, PDF documents from the EOT 2008 Archive. The second step indexed the resulting dataset along several PDF format-specific elements and additional elements of interest. In the third step, queries were formulated to characterize the PDF files contained in the dataset. (NOTE: Technical references at the end of the paper identify the URLs of the maintenance organizations for tools and standards included in this section.)

Processing the Data

The EOT 2008 Archive is comprised of 160,211,356 URI's captured during a seven month period from August 2008 to March 2009 by the partners in the EOT Project. All harvested content was stored in either the ISO standard WARC format or the legacy ARC format. 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. Any partner interested in maintaining a local copy of the Archive could acquire the dataset using the BagIt format and the *rsync* incremental file transfer utility. The UNT Libraries acquired the dataset in the summer of 2009.

Subsequently, the EOT 2008 Archive dataset was groomed for ingestion into the UNT Libraries Digital Collections. One copy was ingested into the UNT Digital Archive system, a preservation repository. A second copy was staged on public-facing servers for access, and an instance of the Open Wayback Machine provided user services. In conjunction with implementing the Open Wayback Machine instance, a comprehensive CDX file was created to provide access to the Archive.

A CDX file contains information about the URLs present in a Web archive and acts as an index to a group of WARC/ARC files. A typical CDX file entry for a URL contains nine fields separated by a whitespace character. Table 3 defines the fields in the EOT2008 Archive CDX files.

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.

Table 3. Fields and Values in a Typical CDX File Entry

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

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, that contained the required eight files. Here is an example:

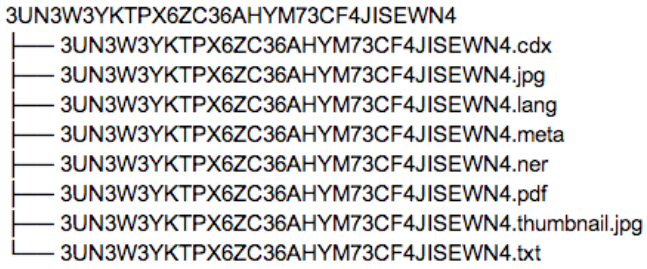

Figure 1. *Completed PDF Sample*

After each of the 4.5 million PDF documents were processed, they were classified as either complete or incomplete samples (Table 4). The distinction was based on whether a PDF file was corrupted or not.

Table 4. PDF Samples (*N* = 4,544,465)

Sample Classification	#	%
Complete	4,404,048	97%
Incomplete	140,417	3%

The incomplete samples represent 3.18% of the dataset and include situations in which the PDF files were corrupt on the original server, corrupted during the harvesting process, or corrupted during the writing of files to the archive WARC/ARC container files.

Indexing the Dataset

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 5). 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.

Table 5. Data Fields Extracted & Indexed from the PDF Samples

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	

Query Formulation

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.

Findings

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 (e.g., .gov and .mil) in the Archive. 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 6 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 6.)

Table 6. Top Level Subdomains by PDF Documents and Pages

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

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 7 shows the distribution of documents by the range of pages per document, from 1 to over 1,001 pages.

Table 7. Distribution of PDF Documents by Range of 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%

PDF Format

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.

Versions

The PDF dataset within the EOT 2008 Archive includes examples of each version, from 1.0 to 1.7. Table 8 identifies the distribution of these PDF versions along with the date the version was initially released.

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.

Table 8. PDF Documents by PDF Version & Release Date

Version	Release Date	#	%	Cumm %
0.0*	-	12	0.00%	0.00%
1.0	1993	14,693	0.33%	0.33%
1.1	1994	69,503	1.58%	1.91%
1.2	1996	807,300	18.33%	20.24%
1.3	2000	1,127,399	25.60%	45.84%
1.4	2001	1,449,508	32.91%	78.76%
1.5	2003	525,647	11.94%	90.69%
1.6	2005	400,174	9.09%	99.78%
1.7	2006	9,812	0.22%	100.00%

* Refers to PDF documents that had internal metadata designating them to be version 0.0.

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. Another example is that there are 20,994 documents that have 1910 listed 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.

Table 9. Encryption Settings for Encrypted PDF Documents

Encryption Settings	#
yes (print:yes copy:yes change:no addNotes:yes)	101,148
yes (print:yes copy:yes change:no addNotes:no)	56,266
yes (print:yes copy:no change:no addNotes:no)	30,387
yes (print:yes copy:yes change:yes addNotes:yes)	12,515
yes (print:yes copy:no change:no addNotes:yes)	3,292
yes (print:yes copy:no change:yes addNotes:yes)	1,107
yes (print:no copy:no change:no addNotes:no)	797
yes (print:yes copy:yes change:yes addNotes:no)	553
yes (print:yes copy:no change:yes addNotes:no)	457
yes (print:no copy:yes change:no addNotes:no)	58
yes (print:no copy:yes change:yes addNotes:yes)	22
yes (print:no copy:no change:yes addNotes:yes)	8
yes (print:no copy:no change:no addNotes:yes)	7
yes (print:no copy:yes change:no addNotes:yes)	7
yes (print:no copy:no change:yes addNotes:no)	2

Discussion

Domains, Subdomains, and Document Size

The .gov Top Level Domain (TLD) is restricted for use by U.S. government entities and agencies. Although this was the dominant domain captured in the EOT 2008 Web Archive, it was not the exclusive domain. Web-published U.S. government information is not bounded by the .gov domain, and the PDF dataset reflects that. This is consistent with the experience of national libraries and archives that seek to capture and archive

their national government publications, many of which are hosted on domains other than country code Top Level Domains (TLDs-cc) [6].

Considering its mandate to publish and disseminate official and authentic U. S. government publications, it is not surprising that the U.S. Government Printing Office (GPO) subdomain (gpo.gov) hosted the largest number of PDF documents, regardless of their page count. However, two agencies that did not make the top six PDF publishers, in terms of total number of PDF documents hosted, emerged as the second and third ranked publishers of PDF documents comprised of 20 or more pages: *gao.gov*, the U.S. Government Accountability Office and *epa.gov*, the U.S. Environmental Protection Agency.

Considering that finding, as well as the fact that 14% of the PDF documents in the dataset are 15 pages or more in length, it may be that document size is an indicator of substantial government publications that might be of interest to a range of information professionals and researchers. It would be worthwhile to determine if these multi-page documents are also hosted by or linked from any GPO subdomains. Are these document titles within or without the GPO's official catalog of publications? Government information professionals in federal depository libraries, as well as the GPO itself, are always vigilant to the discovery of the "fugitive" publications of federal agencies.

PDF Format

While it is an admirable goal of the PDF standard to maintain backwards compatibility with previous versions of the specification, that goal has never been fully achieved. As a consequence, older PDF file formats may be limited in their functionality or impossible to render when opened with later PDF versions. This is one aspect of the problem of obsolescence that preservation repositories handle, generally through format migration or emulation. It is of value to be able to identify documents published with older PDF versions so that curators can estimate their long-term value and can establish appropriate migration and emulation strategies.

The PDF format allows for creators to optimize their documents so that they may be viewed more quickly on the Web. Generally, PDF optimization involves adjusting the ordering of information within the file to allow for linear loading of the files by applications such as Web browsers. Some PDF software adds functionality to the optimization step that can involve compression of image files and other processes that may result in a lower quality file. In the light of long term preservation and access to the highest quality files, the fact that most of the files in this archive are not optimized is promising because optional processes that *might* lower the quality of the files were not implemented.

Only 5% of the documents in the PDF dataset were encrypted. Even though the overall percentage is comparatively low, it represents over 200,000 documents that have varying levels of restriction applied to them. The way in which individual software packages implement these restrictions may have an effect on the discovery or utility of these files in the future.

Embedded Metadata

Metadata extraction from common file formats like PDF is relatively easy with modern tools and libraries. However, the data values held in the internal files often vary widely in their quality. Processes to sanity check some of the values before utilizing them may be needed. One example is the PDF embedded metadata element "creation date".

Because creation dates can be established by different parties (i.e., original creators of the intellectual or artistic content and creators of the PDF version of the content) as well as by the PDF application itself, there will be ambiguity and uncertainty regarding exactly to what they pertain. This suggests caution in several areas, for example, how an archive provider presents users with date information for discrete documents or uses the information to suggest relationships or similarities among content.

On the other hand, the discovery that PDF documents with first page surface areas measuring over 500 square inches typically represented maps, posters, or charts is an interesting and potentially useful finding. It may be that this data element will enable search tools to target discovery of this class of content within a Web archive. Researchers seeking to identify this class of material would be well served with this ability.

Future Work

This analysis of the EOT 2008 Archive's PDF content represents a small step in understanding large bodies of PDF content stored in Web archives. The researchers intend to conduct further analysis in order to take advantage of additional data points, and hopefully to develop a user-friendly interface for querying the dataset.

One area the researchers found particularly interesting was the utilization of metadata fields in the PDF dataset that are typically thought of as descriptive metadata fields. These include title, subject, and author. A full analysis of this data was not addressed in this research but an example pertaining to PDF document "subjects" demonstrates some of the problems. Consider the three subjects that occur most frequently in the PDF documents hosted in the gpo.gov subdomain (Table 10.)

Table 10. Subjects in gpo.gov Subdomain

Subject	# (*N* = 1,082,735)	%
"Extracted Pages"	863,870	79.79%
"" *	219,254	20.25%
"105th Congress"	199	0.02%

* blank value

These numbers suggest that creators of PDF documents in the federal government (specifically the Government Printing Office) do not take advantage of format-supplied metadata fields for embedding descriptive information. Instead they may rely on external metadata catalogs for description and discovery of items.

Additional areas of further exploration might include: (a) the extraction of embedded XMP metadata; (b) name entity information extracted using the Stanford NER tool; (c) color information from the first page of the PDF documents; (d) the entropy of the first page based on histogram information; (e) the

full-text of the document; and (f) the Web graph created by the inlinks within the Archive's content to the PDF documents.

Lastly, the longest amount of time spent in the project workflow was the machine time to process the files. In the future, the research team plans to investigate the use of an infrastructure such as Hadoop to improve process time.

Conclusion

This investigation asked: Is it feasible to describe the content of Web archives by format-specific features? The answer is: "Yes, it is". Further, it seems feasible to take advantage of the findings from this study and use them 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

Because Web archives continue to expand in size and scope, and the time intervals continue to elongate between content capture and Archive query, it is important for libraries and archives to develop new tools that allow users to discover and investigate these rich information repositories. Investigations like the analysis described in this paper can be performed for any content type in Web archives of any size. The ability to provide new and engaging tools creates new avenues for users to discover the "hidden" content in Web archives.

References

[1] Library of Congress, "Web-at-Risk Project". Retrieved January 31, 2013 from http://www.digitalpreservation.gov/partners/webatrisk.html

[2] I. K. Hsieh & K. Murray, "Needs Assessment Survey Report: Abbreviated Version", UNT Digital Library. http://digital.library.unt.edu/ark:/67531/metadc36323/ Accessed January 31, 2013.

[3] University of North Texas Libraries, "eotcd". Retrieved January 31, 2013 from http://research.library.unt.edu/eotcd/wiki/Main_Page

[4] J. Gavin & A. Grotke, "Library Partnership Preserves End-of-Term Government Web Sites", August 14, 2008. Retrieved January 31, 2013 from http://www.loc.gov/today/pr/2008/08-139.html

[5] K. Murray, L. Ko, & M. Phillips, Curation of the End-of-Term Web Archive, Archiving 2012 Final Program and Proceedings, pp. 71-76. (2011).

[6] M. Ras & S. van Bussel, "Web archiving user survey", 2007. Retrieved January 18, 2013 from http://www.kb.nl/sites/default/files/docs/KB_UserSurvey_Webarchive_EN.pdf

Technical References

ImageMagick: http://www.imagemagick.org/

Language-Detection: http://code.google.com/p/language-detection/

PDF ISO 32000-1:2008: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51502

PDF Version Information: http://en.wikipedia.org/wiki/Portable_Document_Format

Solr: http://lucene.apache.org/solr/

Stanford NER: http://nlp.stanford.edu/software/CRF-NER.shtml

xPDF: http://www.foolabs.com/xpdf/

Subdomain References

army.mil	United States Army
bea.gov	U.S. Bureau of Economic Analysis
census.gov	United States Census Bureau
epa.gov	U.S. Environmental Protection Agency
gao.gov	U.S Government Accountability Office
gpo.gov	Government Printing Office
house.gov	The United States House of Representatives
noaa.gov	National Oceanic and Atmospheric Administration
uscis.gov	U.S. Citizenship and Immigration Services
uscourts.gov	United States Courts
usda.gov	U.S. Department of Agriculture

Author Biography

Mark Phillips, Assistant Dean for Digital Libraries, UNT Libraries has been involved with all stages of digital projects at the UNT Libraries for the past ten years. His division is responsible for the management of over 400,000 unique digital items in the UNT Digital Library and The Portal to Texas History. His experience includes the subjects of digital preservation, born digital collection development, digitization, and Web archiving. He acts as the architect for digital library systems at the UNT Libraries.

Kathleen Murray is a postdoctoral research associate at the University of North Texas (UNT) Libraries. Her research primarily involves user studies in the areas of digital libraries and Web archives. She is project manager and a principal researcher for the Classification of the End-of-Term Web Archive project, a three-year research project funded by the Institute of Museum and Library Services.

.