THE CHESAPEAKE PROJECT

Legal Information Archive

First-Year Pilot Project Evaluation

April 2008

Legal Information Archive: The Chesapeake Project

Georgetown University Law Library
Maryland State Law Library
Virginia State Law Library

First-Year Pilot Project Evaluation
April 2008

Contents

I. Executive Summary

The Chesapeake Project is a collaborative, two-year pilot program with the goal of preserving born-digital legal information published directly to the Web. It was implemented in early 2007 under the auspices of the Legal Information Preservation Alliance (LIPA), an independent organization of law libraries supported by the American Association of Law Libraries (AALL), by three LIPA-member libraries: the Georgetown Law Library and the State Law Libraries of Maryland and Virginia.

The following document comprises an evaluation and account of The Chesapeake Project's accomplishments during its inaugural year, spanning from February 27, 2007, to February 29, 2008. During this time, the project digital archive has been populated with more than 2,700 digital items representing nearly 1,300 Web-published titles, the vast majority of which have no print counterpart. Each of these titles were harvested from the Web, stored within a secure digital archive, assigned permanent archive URLs, and cataloged locally and in the WorldCat global catalog, a content network shared by more than 10,000 libraries throughout the world. Today, each archived digital title remains accessible to users through open-access channels and via stable URLs, despite whether or not the original digital files have been altered or removed from their original URLs. More than eight percent of the titles archived between March 2007 and March 2008 by libraries participating in The Chesapeake Project have already disappeared from their original locations on the Web but remain accessible thanks to the project's efforts. Undoubtedly, this figure will increase over time.

Through its first year, The Chesapeake Project has developed a project management model that has accommodated the needs, preservation priorities, and resources of three very different libraries with staffs ranging in size from five to nearly 70. Only one library participating in the project, the Georgetown Law Library, hired a new staff member for the specific purpose of managing the project and coordinating project participants; the other two participating libraries were able to actively contribute to the archive and incorporate digital preservation into the workflows of existing staff librarians. A flexible project collection plan was developed, which provided metadata entry standards for the digital archive while also allowing for flexibility in the development of each participating library's digital archive collection. All three libraries participating in The Chesapeake Project are pleased with the project's progress throughout its first year and enthusiastic about the prospect of continuing the project beyond its pilot phase.

The following represents a snapshot of the findings of The Chesapeake Project's First-Year Evaluation:

Archiving Activity and Numbers

- 2,705 digital items were harvested and archived, along with their respective preservation metadata, in the OCLC Digital Archive, representing roughly 1,266 titles.

- Monthly archiving activity levels varied from 71 to 377 items archived per month. The mean number of items archived per month was 225.42 with a standard deviation of 96.53. The median number of items added per month was 221.

Access Statistics

- Items in The Chesapeake Project's digital archive were accessed total of 5,317 times. Public (non-authenticated) users were responsible for 2,528 instances of access (47%). Libraries participating in The Chesapeake Project accessed their own materials as authenticated users as total of 2,267 times (43%). Authenticated libraries and institutions not participating in the project accessed archived items a total of 533 times (10%).

- Total monthly access figures ranged from a low of 206 instances of access in July 2007 to a high of 979 instances of access in September 2007.

Original URL Inactivity Levels

- Through a sample analysis of the original URLs of titles harvested from the Web and archived through February 29, 2008, it was determined that 8.3 percent of the original URLs had become inactive by March 2008.

- More than 90 percent of the top-level domains in the sample were state (state.[state code].us), organization (.org), and government (.gov) URLs, which represented approximately 41 percent, 32 percent, and 17 percent of the sample, respectively. 10.8 percent of state URLs, 10 percent of government URLs, and 8.3 percent of organization URLs were found to be inactive.

- More than 95 percent of the titles in the sample were in PDF format. Of these, 8.2 percent were found to have inactive original URLs. Four percent of the titles in the sample were in X/HTML formats; these items were found to have a similar inactivity rate of 8.7 percent.

Staffing and Time Commitments

- Staffing and time commitments varied according to institution size, with hours per week devoted to the project ranging by institution from 5 to 30. Participants agree that weekly hours have decreased as they have become more familiar with project tools and resource discovery and selection methods.

- Cataloging archived items was ranked as the most time-consuming project-related activity, followed, in decreasing order, by harvesting and archiving digital materials, selecting and monitoring Web-based publications for archiving, and general project coordination.

Challenges and Problems Encountered

- The greatest challenge posed to The Chesapeake Project in its first year was the transition from the original OCLC Digital Archive to the new CONTENTdm/dark digital archive system, which is ongoing at the present time.

- Additional challenges cited by project participants had to do with 1) the selection and management of content placed in the digital archive; 2) procedures, policies, and project standard development; 3) learning to use OCLC systems, as well as learning to do original cataloging; 4) the loss of an influential project leader.

Progress toward the Realization of Project Mission and Vision

- Participants agree that although great strides have been made in the first year, in order to truly realize the project's mission and vision, a concerted effort to raise awareness and educate others about the work of The Chesapeake Project will be instrumental in the project's second and final year.

II. Introduction and Background

Introduction

Access to information is a fundamental right of a democratic citizenry. For our democracy to thrive, the general public, lawmakers, law practitioners, and law scholars alike must have access to the legal information that describes, explains, critiques, and comprises our laws, systems of justice, and legal rights.

As citizens of the 21st century, we have not only witnessed the emergence of information created in digital formats and disseminated via electronic channels, but we have also seen popular and widespread adoption of digital culture. Although information in digital formats is compact, easily transportable, and instantly accessible, it is also at an alarming risk for permanent loss. Digital formats are threatened by obsolescence as new technology replaces existing systems and applications, and the current lifespan of digital media is uncertain.

Of particular concern is the ephemerality of legal information published directly on the free Web, without a printed counterpart. Access to these materials, which are often independently posted online by government entities, agencies, scholarly societies, and other organizations, can be unexpectedly and permanently lost as files are removed and URLs are changed or inactivated through routine and seemingly innocuous Web site maintenance activities.

In March 2007, the Georgetown University Law Library and the State Law Libraries of Maryland and Virginia embarked upon a collaborative pilot project to address the challenge of preserving legal information published online. This project, The Chesapeake Project, has been initiated under the auspices of the Legal Information Preservation Alliance, or LIPA, an organization that formed in 2003 to support, guide, and advance a national strategy for the preservation of legal information in various formats. Ultimately, at the close of its two-year pilot phase, The Chesapeake Project aspires to help fulfill LIPA's mission by either instigating or evolving into an organized, nationwide digital preservation program for legal information, a legal information archive.

To document and track the progress of The Chesapeake Project through the course of its pilot phase, and to provide a point of reference for future project improvement and benchmark-setting, a series of annual evaluations are being conducted at the project's first- and second-year marks. The following document represents a report of The Chesapeake Project's first-year evaluation, from March 2007–March 2008.

Overview of The Chesapeake Project

The Chesapeake Project is a two-year pilot digital preservation program established to preserve and ensure permanent access to vital legal information currently available in digital formats on the World Wide Web. Although preliminary planning for the project began in 2006, the project collection and archiving activities by the three collaborating libraries — the Georgetown University Law Library and the State Law Libraries of Maryland and Virginia — began in March 2007.

Users

The Chesapeake Project serves patrons of the Georgetown University Law Library and the State Law Libraries of Maryland and Virginia as its primary user group. This patron group consists of law practitioners, law faculty members, law students, justices and their staff members, judges and their staff members, and state government officials and their staff members. Secondary users include law scholars and students not affiliated with participating libraries, as well as the general public.

Collections

Because the World Wide Web is an ever-expanding source of digital resources and publications, delimiting the collection scope for the purpose of the pilot was a crucial first step at the project's outset. Each library participating in The Chesapeake Project defined its own initial scope of Web-harvesting and digital-archive collection

development priorities, contextualized within the larger collection theme of legal information published online, based on institutional missions and mandates, and also considering the needs and research interests of its users.

The digital-archive collection of the Georgetown University Law Library is a largely thematic collection of secondary legal sources, selected based on research and educational areas of interest at the Law Center, including legal resources in the areas of animal law, journalism law, copyright law, law and public health, environmental law and policy, conflict resolution, human rights, and the Supreme Court. Additional collection areas include law-related publications produced by and about the District of Columbia, as well as select high-interest reports and studies produced by federal commissions.

The Maryland State Law Library's digital-archive collection consists of selected digital materials that describe, analyze, document, propose, clarify, or define public-policy and legal issues that affect the citizens of the state of Maryland. Of particular interest are task force reports mandated by the Maryland General Assembly, reports of gubernatorial commissions, publications issued by the Maryland Judiciary, and major reports issued by Maryland executive agencies. The Law Library collects selected publications from Maryland community and research organizations whose studies and reports provide an analysis of major issues of public policy and law.

The digital-archive collection of the Virginia State Law Library consists of all publications issued online by the Supreme Court of Virginia, as well as publications issued by the Judicial Council of Virginia and the range of administrative divisions, commissions, and task forces operating within Virginia's judicial branch of government.

Preservation & Access System

For the storage, preservation, and management of these digital collections, libraries participating in The Chesapeake Project chose to utilize the OCLC Digital Archive, which adheres to the ISO reference model for an Open Archival Information System (OAIS). Access and discovery of archived items is made possible by a unique OCLC Digital Archive URL, a permanent URL with OpenURL syntax, which is generated for each object in The Chesapeake Project's collections. The OCLC Digital Archive URL is added to local records as well as OCLC bibliographic records, providing direct access to archived items via records in participating libraries' local OPACs, as well as OCLC' s WorldCat and FirstSearch.

In April 2008, shortly after the project's the first-year mark, OCLC will transition The Chesapeake Project's archived collections and metadata from the original OCLC Digital Archive to an enhanced, two-tiered digital-preservation and access system, which will manage and provide access to archived collections via CONTENTdm, while preserving master files within a separate, dark digital archive. All existing OCLC Digital Archive URLs will resolve to the files' new locations within CONTENTdm. In May 2008, following the full transition of files and metadata to the new CONTENTdm-based system, the current OCLC Digital Archive system will be deactivated.

III. Project Evaluation Overview

Although the ambitions and vision of The Chesapeake Project are bold, it cannot be forgotten that The Chesapeake Project is, first and foremost, a pilot project. It has been established as a trial run-through to investigate the feasibility of establishing a collaborative, national digital archive for legal information. As such, the project was not created with a predetermined set of benchmarks and objectives against which to measure its progress; rather, project participants have utilized the first year of the pilot to familiarize themselves with the digital archiving process, create shared documentation to guide project participation, assess digital-archiving costs and necessary staffing commitments, and develop reasonable expectations for progress in digital archiving and archive collection development.

While evaluations of individual digital libraries and comparative assessments of archiving software options can be found in the professional literature, relatively little literature can be found on the topic of evaluating digital-archiving and Web-harvesting programs. In June 2007, project participants convened to discuss, among other agenda items, project evaluation parameters for inclusion in The Chesapeake Project Collection Plan document. At that time, it was recommended that the project's first-year evaluation include the following points of discussion:

- Description of the project and archive;
- Count of archived items and titles;
- Archived-item access statistics;
- Test sample to determine percentage of archived items altered or removed from the Web;
- Progress toward accomplishment of project mission and vision;
- Challenges encountered; and
- Recommendations for project improvement.

An additional point of discussion added to this list in January 2008 is a description of staffing and time commitments required for participation in the project. Each of these points of discussion is integrated into the quantitative evaluation and qualitative evaluation sections that follow.

IV. Quantitative Evaluation

A. Archived-Item Count and Access Statistics

1. Introduction

Research Objectives

In an effort to quantifiably evaluate the progress of The Chesapeake Project through its inaugural year, this section of the evaluation provides an account and discussion of the number of items archived by The Chesapeake Project as well as access statistics for archived items from March 2007–March 2008.

This study aims to answer the following questions:

- How many items were added to the digital archive during The Chesapeake Project's first year?
- How many archived items were accessed by users during The Chesapeake Project's first year?

Definitions

The term *archived items* refers to the number of discrete information packages or publication issues that are first harvested, or downloaded, from the Internet and then ingested, or stored, into the OCLC Digital Archive. Each archived item has a single, corresponding preservation metadata record in the digital archive. It is important to note that the count of archived items differs from that of archived titles. The term *archived titles* refers to the number of individual monograph or serial publications harvested from the Internet and ingested into the digital archive. Each archived title has a single, corresponding bibliographic record in OCLC's global WorldCat catalog. As in the case of multi-part monographs or serial Web publications, it is not uncommon for a single archived title to be comprised of more than one archived item.

Access figures are a measure of usage describing the number of times archived items are viewed online by users who connect to the items via OCLC Digital Archive URLs. For the purpose of this analysis, access figures are delineated as either public access figures or authenticated access figures. *Public access figures* log the number of times that non-authenticated users, those who are not logged in to the OCLC system as an OCLC-member library or institution, have viewed archived items as general users online. *Authenticated access figures* log the number of times that users have logged in to an OCLC system and viewed archived items as an authenticated OCLC-member library or institution.

In the following analysis, authenticated access figures have been further divided into two final categories: project-participant access figures and authenticated non-participant access figures. *Project-participant access figures* log the number of times that the three libraries participating in The Chesapeake Project have logged in to an OCLC system and viewed archived items. *Authenticated non-participant access figures* log the number of times that libraries and institutions who are not among the three libraries participating in The Chesapeake Project have logged in to an OCLC system and viewed archived items. Segregating these two sets of authenticated users is important for the purpose of determining the extent to which participating libraries are accessing their own archived items through the course of testing archive URLs and cataloging archived titles. Also of interest is the extent to which other OCLC-member libraries and institutions are making use of archived items during research, reference, and cataloging activities.

2. Methodology

Data Gathering

Quantitative data were gathered from two separate sources: the digital archive itself and OCLC-generated reports. The archived item count was obtained through a manual count of items in the OCLC Digital Archive as of March 6, 2008. This count comprises 2,705 archived items ingested into the digital archive between the dates of February 27, 2007, and February 29, 2008, by libraries participating in The Chesapeake Project. In addition to the total item count, items added to the archive by month were also counted. These total and monthly counts of archived items were made possible by an "ingested on" date which is automatically generated for each item metadata record upon ingest.

Access figures were obtained from OCLC Product Services, which provides monthly reports of access summaries that log authenticated OCLC-member institutions accessing items from the digital archive, including those libraries participating in The Chesapeake Project, as well as a count of items accessed by non-authenticated "public" users.

Access statistics provided by OCLC Product Services are transaction logs that differentiate and list which authenticated OCLC-member institutions accessed digital archive items, while also including a count of how many archived items have been accessed, in a given month, by each authenticated institution. However, authenticated access figures and public access figures do not necessarily provide a definitive count of institution use of archived items against public or patron use of archived items. This possible discrepancy can occur when an OCLC-member institution, including those libraries participating in The Chesapeake Project, accesses a digital archive item without first logging into an OCLC system. In such a situation, the access instance would be added to the public access count, rather than to the authenticated access count.

An additional threat to the validity of access statistics provided by OCLC Product Services is the method by which access statistics are gathered, by tallying *items* accessed as opposed to *titles* accessed. In the instance of multi-part serials and monographs, archived titles that are made up of multiple archived items, the access count can be inflated. For example, if a user accesses a single serial title comprising 25 archived items, the access count will increase by 25, instead of by one.

Data Analysis

To achieve a comprehensive view of the data collected, data were charted and analyzed from various vantage points. The following list delineates how data were charted and analyzed to provide varying perspectives on The Chesapeake Project's first-year archiving and access activities:

1) *Archived Items:* The count of archived items is presented in the form of descriptive statistics for the purpose of tracking and anticipating archiving activity levels. Archived item statistics are charted a) cumulatively, tracking the sum total number of items in the archive at the end of each month; and b) by activity level, tracking only the number of items added to the archive by month, rather than the cumulative total. To gauge the level of monthly archiving activity during the project's first year, the mean, median, and standard deviation are calculated for the non-cumulative, monthly activity figures.

2) *Access figures:* Access figures are also put forth as descriptive statistics for the purpose of tracking and anticipating access statistics of various types of users. Access figures presented include total access figures, public access figures, project-participant access figures, and authenticated non-participant access figures. Each of these sets of access figures are charted a) cumulatively, tracking the sum total of access numbers at the end of each month; and b) by activity level, tracking only the number of archived items accessed by month, rather than the cumulative total.

3. Results

Archived Items

In the first year of The Chesapeake Project, a total cumulative number of 2,705 items were harvested from the Web and archived by the three libraries participating in the project.

Monthly archiving activity levels varied, ranging from a low of 71 items archived by project participants in February 2008, to a high of 377 items archived in August 2007. See Table 1 for a side-by-side listing of items added by month against the sum total of archived items. Figure 1 provides a graphical representation of the first-year cumulative archived-item count.

The mean number of items added to the digital archive per month was 225.42 with a standard deviation of 96.53. The median number of items added to the digital archive per month was 221. Monthly archiving activity levels are charted separately in Figure 2.

MONTH	ITEMS ADDED	TOTAL ITEMS
Mar-07	115	115
Apr-07	334	449
May-07	169	618
Jun-07	234	852
Jul-07	145	997
Aug-07	377	1374
Sep-07	236	1610
Oct-07	356	1966
Nov-07	313	2279
Dec-07	208	2487
Jan-08	147	2634
Feb-08	71	**2705**

Table 1. The Chesapeake Project first-year monthly archiving activity and cumulative items added to the digital archive

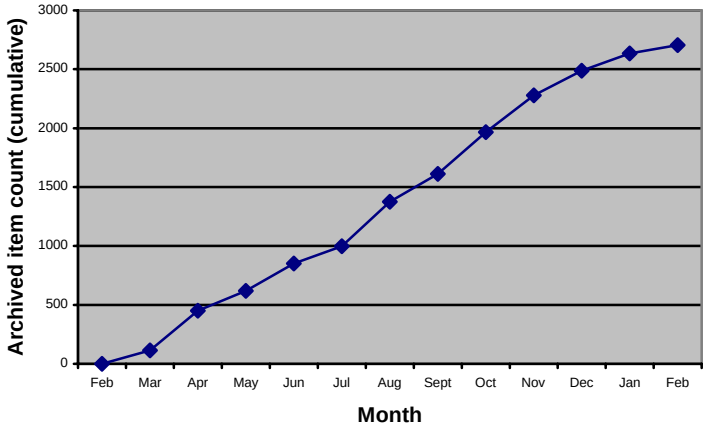

Figure 1. The Chesapeake Project first-year cumulative total of archived items

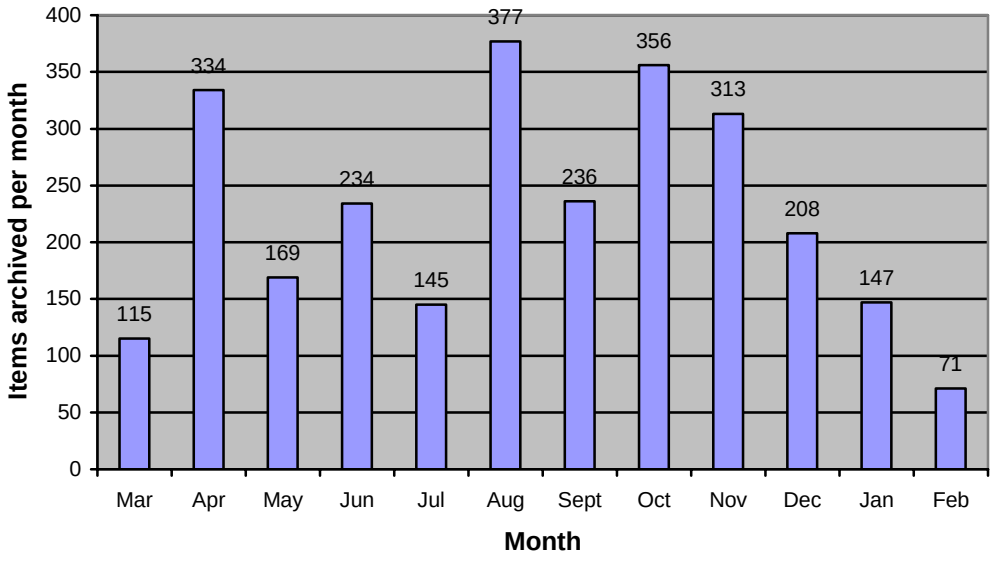

Figure 2. Number of new items added to The Chesapeake Project digital archive by month

Access Figures

Cumulative Access Figures

During the first year of The Chesapeake Project, archived items were accessed total cumulative number of 5,317 times (including public and authenticated access figures). Of this total cumulative access figure, archived items were accessed a sum total of 2,528 times by public (non-authenticated) users, 2,267 times by authenticated libraries participating in The Chesapeake Project, and 522 times by authenticated non-participant institutions. Of the total number of times archived items were accessed over The Chesapeake Project's first year, about 47 percent comprises public access figures, 43 percent comprises project-participant access figures, and 10 percent comprises authenticated non-participant access figures. Cumulative access figures by access type are available in Table 2; the division of total cumulative access figures is illustrated by Figures 3 and 4.

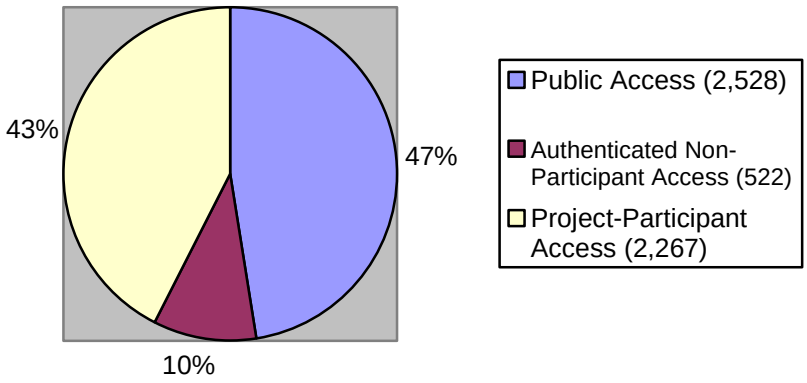

Figure 3. Cumulative total annual access figure of 5,317, divided by access type: public (47%), project-participant (43%), and authenticated non-participant (10%) access

MONTH	CUMULATIVE TOTAL ACCESS	CUMULATIVE PUBLIC ACCESS	CUM. PROJECT-PARTICIPANT ACCESS	CUM. AUTH. NON-PARTICIPANT ACCESS
Mar-07	277	85	188	4
Apr-07	776	284	484	8
May-07	1152	459	671	22
Jun-07	1379	580	767	32
Jul-07	1585	666	866	53
Aug-07	1926	710	1124	92
Sep-07	2905	1481	1276	148
Oct-07	3573	1871	1494	208
Nov-07	4213	2224	1699	290
Dec-07	4506	2304	1870	332
Jan-08	4820	2431	1990	399
Feb-08	**5317**	**2528**	**2267**	**522**

Table 2. The Chesapeake Project first-year cumulative digital archive access figures

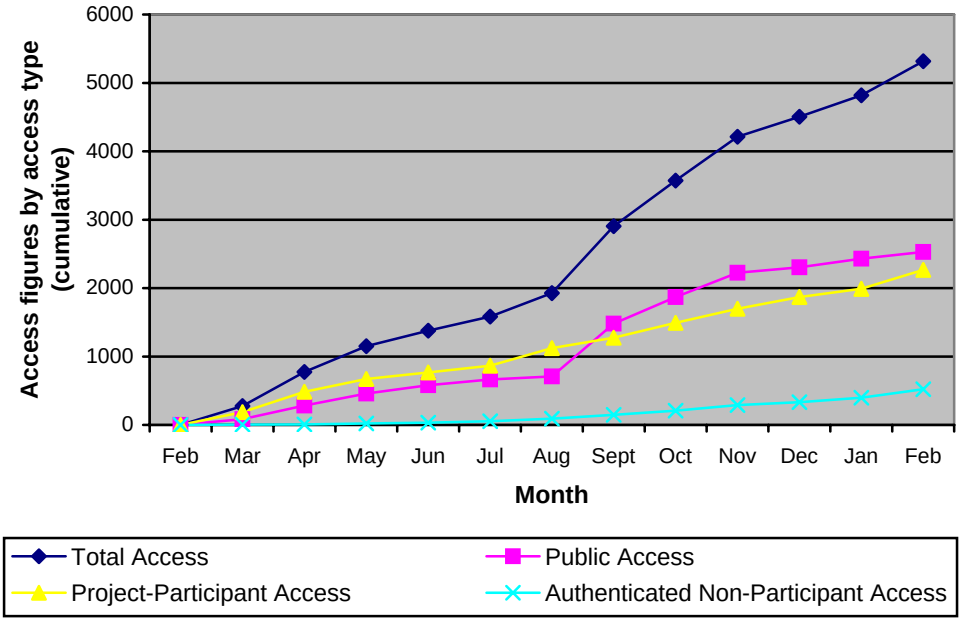

Figure 4. Cumulative access figures by month, presented by types of access (total access, public access, project-participant access, and authenticated non-participant access)

Monthly Access Figures

Access activity levels varied both by month and by type of access. Total monthly access figures ranged from a low of 206 instances of access in July 2007, to a high of 979 instances of access in September 2007. The mean number of times items were accessed from the digital archive per month was 443.01 with a standard deviation of 217.41. The median number of times items were accessed from the digital archive per month was 358.5.

Monthly public access figures ranged from a low of 44 instances of access in August 2007, to a high of 771 instances of access in September 2007. The mean number of times items were publicly accessed from the digital archive per month was 210.67 with a standard deviation of 198.32. The median number of times items were accessed from the digital archive per month was 124.

Monthly project-participant access figures ranged from a low of 96 instances of access in June 2007, to a high of 296 instances of access in April 2007. The mean number of times items were accessed per month by authenticated project-participants was 188.92 with a standard deviation of 63.42. The median number of times items were accessed from the digital archive by authenticated project-participants per month was 187.5.

Monthly authenticated non-participant access figures ranged from a low of 4 instances of access in March and April 2007, to a high of 123 instances of access in February 2008. The mean number of times items were accessed per month by authenticated non-participants was 43.5 with a standard deviation of 34.67. The median number of times items were accessed from the digital archive by authenticated non-participants was per month was 40.5.

See Table 3 for a listing of monthly access figures by access type. Figure 5 provides a graphical representation of the variance in first-year access figures by month and access type.

MONTH	TOTAL ACCESS by MONTH	PUBLIC ACCESS by MONTH	PROJECT-PARTICIPANT ACCESS by MONTH	AUTHENTICATED NON-PARTICIPANT ACCESS by MONTH
Mar-07	277	85	188	4
Apr-07	499	199	296	4
May-07	376	175	187	14
Jun-07	227	121	96	10
Jul-07	206	86	99	21
Aug-07	341	44	258	39
Sep-07	979	771	152	56
Oct-07	668	390	218	60
Nov-07	640	353	205	82
Dec-07	293	80	171	42
Jan-08	314	127	120	67
Feb-08	497	97	277	123

Table 3. The Chesapeake Project first-year monthly digital archive access figures

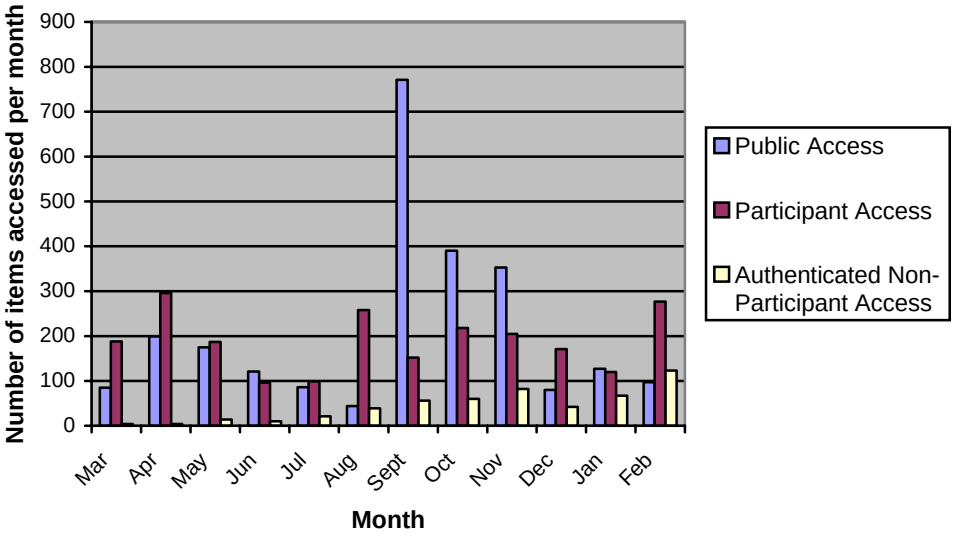

Figure 5. Number of items accessed per month, divided by access type (public access, project-participant access, and authenticated non-participant access)

4. Discussion

Archived Items

A total of 2,705 items were added to the digital archive by the three institutions participating in The Chesapeake Project. During the year, the highest levels of archiving activity (more than 300 items archived during the month) occurred in the months of April, August, October, and November 2007. The high activity in April may be speculatively linked to project participants' increasing familiarity with the digital archiving system and eagerness to begin building collections.

Notably low archiving activity levels occurred in March 2007 and February 2008, with 115 and 71 items archived per month, respectively. As March 2007 represented the start of the pilot project, this low activity level is likely to have been caused by time spent by project participants' learning to use the digital archiving system and developing workflows and metadata-entry policy. Low activity levels in February 2008 may be attributed to the pending transition of archived items from the original OCLC Digital Archive system to the new, two-tiered, CONTENTdm/dark digital archive system; this transition required project participants and digital archive curators to devote time to preparing files for the new system, as opposed to harvesting and archiving new items.

Tracking archiving activity levels over the next year of the pilot will be an interesting exercise that may provide some insight into activity levels that may be predicted if and when the project expands in its post-pilot phase. It is also worth noting that digital archive system transitions, which are inevitable given the ever-progressing nature of technology and digital preservation, will require time to be diverted from archiving activity in order to manage the transition.

Access Figures

At first glance, the first-year access figures, compared to the number of items in the digital archive, are so impressive that they arouse suspicion; of 2,705 items placed into the digital archive during the year, there were 5,317 instances of access. However, once these access figures are deconstructed and analyzed, a clearer and more accurate picture of access statistics can be discerned.

Of the total 5,317 instances of access, 2,267 access instances can be traced directly to authenticated project participant institutions, which are likely to have accessed their own harvested and archived items

to test and verify the quality of the harvest and status of these items during the archiving and cataloging processes. However, it is important to note that authenticated institution-based users not affiliated with the project may also access archived materials in the course of legitimate research and reference duties, which means that this figure cannot always be entirely attributable to active project participants.

2,528 instances of public, or non-authenticated, access were also tracked. While this figure appears to represent general public users, it is important to note that project participants may have accessed archived items (via WorldCat.org, for example) without first being authenticated, thereby inflating this number.

522 instances of access can be irrefutably attributed to authenticated non-project participants, as this figure represents instances of access by named OCLC-member institutions, excluding project participants. It is likely these institutions accessed these items in the course of research activities and adding records with OCLC Digital Archive URLs to their own local catalogs.

B. Archived Titles with Inactive Original URLs

1. Introduction

Research Objectives

The mission of The Chesapeake Project is "to stabilize, preserve, and ensure permanent access to critical born-digital legal materials on the World Wide Web," as born-digital materials published on the free Web are believed to be especially ephemeral and at high risk for loss. Therefore, this section of the evaluation continues the quantitative assessment of The Chesapeake Project by investigating the extent to which the original URLs of titles archived throughout the first year of the pilot project have become inactive. The objective of this project is to examine a statistically significant sample of archived titles in order to answer the following questions:

- What percentage of original URLs are no longer active?
- What are the top-level domains (such as .gov, .com, .org, .us) of original URLs that are no longer active?
- What are the file format types (such as PDF, HTML, or Word Document) of original URLs that are no longer active?

Definitions

The term *archived titles* refers to the number of individual monograph or serial publications harvested from the Internet and ingested into the digital archive. Each archived title has a single, corresponding bibliographic record in OCLC's global WorldCat catalog. It is important to note that multiple archived items may comprise a single archived title, as in the case of multi-part monographs or serial Web publications.

URLs are Uniform Resource Locators, or Internet addresses directing to file sites on the World Wide Web. *Inactive URLs*, for the purpose of this study, are Internet addresses that no longer direct to the files that were originally archived from the locations to which those URLs direct on the World Wide Web. These inactive URLs indicate that the content has been lost, removed, or relocated from its original or previous location on the Web.

Top-level domains are the domain-name suffixes appearing following the final "dot" in a Web site's domain name sequence. Example top-level domains include .gov, .com, .org, and .us; these suffixes can be used to indicate the organization-type of a Web site, such as governmental (.gov), commercial (.com), or educational (.edu).

File format types refer to the digital manifestations of the resources located at each URL. File format types must be compatible with an operating system's platform and software applications in order to render their files' content. Example file format types include X/HTML, PDF, and Word Document files.

2. Methodology

A master list of archived titles, comprising all titles harvested from the Internet and placed into the digital archive from the start of the project, along with each title's corresponding OCLC bibliographic record number, was amassed by project participants at the three collaborating institutions on March 14, 2008. All titles with a digital archive "ingest date" of after February 29, 2008, were removed from the master list, leaving a total of 1,266 titles archived between the dates of February 27, 2007, and February 29, 2008.

From this list of 1,266 titles, a sample of 579 OCLC bibliographic record numbers was randomly selected, ensuring results at a 95 percent confidence level and confidence interval of +/- 3. To assess whether or not the original URL of the archived title remained active or not, the OCLC number was used to retrieve the preservation metadata record from the digital archive. Each metadata record provides the original URL from which the archived item was harvested from the Internet, and these original URLs were checked for inactivity.

A spreadsheet was created for the project on which researchers tracked each sample title's OCLC number, original URL, and whether or not the URL was active. These URL activity checks took place between March 19 and March 28, 2008.

In addition to recording URL activity or inactivity, researchers also tracked the top-level domain and file format type for the files found at each URL. In some cases, the URLs contained more than one file format type, such as both HTML and PDF.

Researchers were given special instructions for assessing serial or multi-part monograph titles. As these titles often require multiple harvests from multiple URLs, and are thus associated with multiple preservation metadata records in the digital archive, researchers were instructed to check the original URL, top-level domain, and format type of the record appearing at the mid-point of the results list only; in other words, neither the earliest nor the most recently harvested record was analyzed.

3. Results

Inactive Original URLs

Forty-eight out of a sample of 579 URLs tested were found to be inactive. Given the total of total of 1,266 titles archived, it can be inferred with a 95 percent confidence level and a confidence interval of +/- 3 that 8.3 percent of the original URLs of all titles harvested and archived during the first year of The Chesapeake Project had become inactive by March 2008.

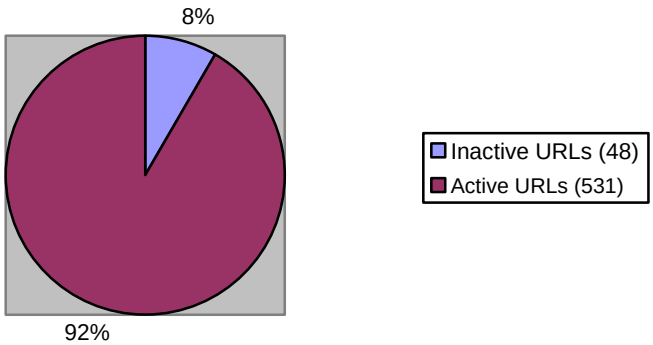

Figure 6. Of a sample of 579 titles harvested from the Internet and archived by The Chesapeake Project between February 27, 2007, and February 29, 2008, the original URLs of 48 archived titles, or 8.3 percent, were found to be inactive by March 2008.

Inactive Original URLs and Top-Level Domains

More than 90 percent of the top-level domains in the sample were state (state.[state code].us), organization (.org), and government (.gov) URLs, which represented approximately 41 percent, 32 percent, and 17 percent of the sample, respectively. Of these three top-level domains, 10.8 percent of state URLs were found to be inactive, 10 percent of government URLs were found to be inactive, and 8.3 percent of the organization URLs were found to be inactive.

Although education (.edu) and commercial (.com) URLs represented a much smaller portion of the sample, both top-level domains were found to have relatively high inactivity levels of 11.8 and 15.4 percent, respectively. A list of all top-level domains found in the sample, along with their inactivity rates, is available in Table 4.

TOP-LEVEL DOMAIN	TOTAL IN SAMPLE	ACTIVE	INACTIVE	PERCENT INACTIVE
.state.__.us	240	214	26	10.8%
.org	184	177	7	8.3%
.gov	100	90	10	10%
.edu	17	15	2	11.8%
.com	13	11	2	15.4%
.net	11	11	0	—
.mil	3	3	0	—
.us	3	3	0	—
.info	2	1	1	50%
.uk	2	2	0	—
.au	1	1	0	—
.ca	1	1	0	—
.int	1	1	0	—
[IP address]	1	1	0	—

Table 4. The top-level domains of active and inactive original URLs for titles in the sample.

Inactive URLs and Format Types

More than 95 percent of the titles in the sample were harvested and archived in PDF format. Of these titles, 8.2 percent were found to have inactive original URLs. A much smaller portion of the sample, 4 percent, was represented by titles in X/HTML formats. Interestingly, these items were found to have a similar inactivity rate of 8.7 percent. Other format types found in the sample included combination HTML/PDF titles and Word Documents. A list of all format found in the sample, along with their inactivity rates, is available in Table 5.

FORMAT TYPE	TOTAL IN SAMPLE	ACTIVE	INACTIVE	PERCENT INACTIVE
PDF	552	507	45	8.2%
X/HTML	23	21	2	8.7%
HTML/PDF	3	2	1	33.3%
Word Doc	1	1	0	—

Table 5. The format types of active and inactive original URLs for titles in the sample.

4. Discussion

It is difficult to determine the typical lifespan of a Web resource. A 2000 report (published via a Web page that has since become inactive, but documented by the Berkeley School of Information Management & Systems) estimated the average lifespan of a Web site to be 44 days, or about six weeks.[1] The non-profit Internet Archive, perhaps best-known for its Wayback Machine, increases this lifespan estimation from 44 up to 75 days.[2] Research of Web citations appearing in scientific journals found in 2003 that 3.8 percent of Web citation URLs had become inactive within 3 months of the citing article's publication, 10 percent of cited URLs had become inactive within 15 months, and 13 percent had become inactive within 27 months.[3] A similar but more recent study of Web citations in New Zealand found that 30 percent of cited URLs, appearing within a small sample of journal articles published in the years 2002–2005, had become inactive.[4]

Our findings of an 8 percent inactivity rate among original URLs for titles harvested and archived during the first year of The Chesapeake Project seems to be average, given these previous findings, particularly considering that every URL in the sample had been active within the past year, including those of some resources in the sample harvested as recently as February 2008.

The majority of titles (slightly more than 90 percent) harvested and placed into the digital archive during the first year of The Chesapeake Project came from state, government, and organization Web sites, as determined by their top-level domains. Of these, state and government Web resources seem to have a slightly higher risk for relocation or removal from the Web, in comparison to organization-published Web resources. More than 95 percent of the titles archived were also published in PDF format, compared to about 4 percent published as X/HTML Web pages. However, in our sample, there appeared to be little significant difference in URL inactivity levels for titles published in PDF or X/HTML formats, with 8.2 and 8.7 percent inactivity rates, respectively.

The sample of URLs studied for this project will be kept on file and re-analyzed, and compared against the current findings as well as against a newer sample drawn at the project's two-year mark.

[1] Retrieved March 12, 2008 from http://www2.sims.berkeley.edu/research/projects/how-much-info/internet/rawdata.html. Cited in Lyman, P. (2002). Archiving the World Wide Web. *Building a national strategy for digital preservation: Issues in media archiving.* Retrieved March 12, 2008 from the Council on Library and Information Resources: http://www.clir.org/PUBS/reports/pub106/pub106.pdf#page=42

[2] See Internet Archive: Wayback Machine at http://www.archive.org/web/web.php.

[3] Devalle, P., et al. (2003). Going, going, gone: Lost Internet references. *Science 302(31),* 787-788. Retrieved Feb. 25, 2008 from Science Magazine: http://www.sciencemag.org/cgi/content/full/302/5646/787

[4] Parker, A. (2007). Link rot: How quality of electronic citations affects the quality of New Zealand scholarly literature. Retrieved March 9, 2008, from Coda, An Institutional Repository for the New Zealand ITP Sector: http://www.coda.ac.nz/whitireia_library_jo/1/

V. Qualitative Evaluation

Introduction

To provide a qualitative evaluation of The Chesapeake Project at its first-year mark, project participants sought to analyze 1) the staffing and time committed by each participating institution through the project's first year; 2) the challenges and problems encountered through the project's first year, and the success to which these challenges were addressed and resolved; and 3) the extent to which the project had progressed toward the accomplishment of its mission and vision.

This qualitative evaluation required the input and contribution of project participants at all levels. To gauge participants' views and allow for the contribution of all project members to the evaluation, a Web-based survey instrument was developed and distributed to project participants. The survey was divided into four parts, which consisted of primarily open-ended questions, with some multiple-choice and Likert items. The first part provided the project's mission and vision statements and asked for commentary on and an evaluation of the extent to which the project's mission and vision had been accomplished. The second part was an assessment of the staffing required and time devoted to the project through the first year. The third part addressed project challenges and resolutions, and the fourth part gave participants an opportunity to comment freely on the project's accomplishments and areas in need of improvement.

A total of six respondents completed and submitted the survey. This number included two law library directors, three project curators/archivists, and one cataloger.

A. Staffing and Time Commitment

Staffing and time devoted to the project varied through the first year by institution, which, as respondents noted, is largely due to differences in library size and responsibility. However, the staffing pattern that appears to emerge among libraries participating in the project is as follows: library leaders serve in an administrative/decision-making role, a staff librarian from each institution is appointed to act as the primary project coordinator, and, depending upon the library's resources, cataloging and technical services librarians provide additional project assistance and guidance.

The Georgetown Law Library has a library staff of nearly 70, and hired a full-time Digital Preservation Librarian whose primary responsibility is to manage The Chesapeake Project, develop project documentation, and archive and catalog titles selected for inclusion in the digital archive. This librarian is designated as the central project coordinator, and as such, facilitates communication as well as the overall working relationship among the three libraries and also acts as the liaison between project participants and their account representatives at OCLC. A team of selectors, comprised of subject specials on the library reference staff, was appointed to support the project by assisting with selection of Web-published items to be archived. Additionally, the Law Library Director, Associate Law Librarian for Collection Services, Head of Cataloging, and Serials/Electronic Collections Librarian serve as project advisors and participate in upper-level project administration, decision-making, and planning.

The Maryland State Law Library has a staff of 15. The primary responsibility for coordinating, curating, and archiving items for The Chesapeake Project is given to the State Publications Librarian. This librarian receives support and guidance from the Library's Head of Technical Services and Head of Electronic Services. The Library Director assists with project planning and strategy.

The Virginia State Law Library has a smaller staff of five, and the Assistant Law Librarian is responsible for archiving and cataloging items selected for inclusion in The Chesapeake Project. The State Law Librarian participates in the project as an administrator, by participating in project meetings and upper-level decision making activities. Due to the small staff size at the Virginia State Law Library, additional staff members cannot be allocated to assist with the project; therefore, time must be carefully managed to ensure that the Assistant Law Librarian is

able to accomplish project-related digital archiving responsibilities while also managing and assisting with other library projects.

Time devoted to the project on a weekly basis by coordinating librarians at the three institutions varied, but almost all participants have reported a reduction in time spent on project-related activities as they become increasingly familiar with the digital archiving tools and as internal work processes are streamlined. Georgetown's project coordinator, whose primary responsibility is project management, devoted approximately 30 hours per week to the project, a figure that is expected to decrease in the coming year. At the Maryland State Law Library, the combined hours spent by two primary librarians on project activities equaled about 12 per week. The Virginia State Law Library project coordinator initially spent about 15 hours per week on the project; at present that figure has dropped to 5 hours per week.

Librarians working on The Chesapeake Project ranked project-related activities from most- to least-time consuming, with the task of cataloging archived items ranking as the most time-consuming project-related activity, due largely to the fact that most items harvested and preserved as part of The Chesapeake Project represent fugitive documents and gray literature, which require original cataloging. Harvesting and archiving digital materials is ranked as the next most time-consuming activity, followed by selecting and monitoring Web-based publications for archiving. The least time-consuming activity was general project coordination. One respondent noted that training was also a time-consuming activity early on in the project.

With the transition to the new CONTENTdm/dark digital archive system, additional time may be required for training during the first few months following the introduction of the new system; however, because the new system integrates the cataloging and archiving processes, time savings may ultimately be realized during the pilot's second year.

B. Challenges and Problems Encountered

The primary challenge encountered during the first year of The Chesapeake Project, which was cited by all project survey respondents, is the current transition from the original OCLC Digital Archive to the new CONTENTdm/dark digital archive system. However, most respondents agreed that this challenge has been adequately addressed through persistent communication and an open dialogue, both with OCLC and between project participants. Project participants were given clear instructions from OCLC and allocated the necessary time and resources to sufficiently prepare for the transition. The new system is priced comparably to the former system (roughly $14,000 for the first year, and $8,500 for the second year, divided between the three participating libraries), and OCLC will honor the original terms of The Chesapeake Project's pilot agreement before moving forward with the new system pricing structure. Because additional libraries may be joining the project once the end of the pilot agreement is reached, ongoing discussions with OCLC are required, as is the development of new policy materials to accommodate the new system.

Additional challenges cited by project participants had to do with the selection and management of content placed in the digital archive, specifically, locating material online, managing the display of multi-part PDF documents, and managing serial publications. Most of these challenges are being addressed internally within individual institutions, through the use of project tools and systems, and through collaboration with and advice from project partners.

Project participants also identified the development of procedures, policies, and project standards as a challenge; although institutions must address issues such as division of duties internally, project participants have thus far successfully developed shared documentation for metadata entry, as well as cataloging procedures to ensure that project records are consistent, comprehensive, and contain information deemed important for the purpose of the project, such as notes about the "provenance," or original URL, and the file format and version of each harvested title. Policy development will be an ongoing and evolving task.

Learning to use OCLC systems was cited as a major challenge. These systems include not only the OCLC Digital Archive and CONTENTdm systems, but also the OCLC Connexion Browser, Connexion Client, and Connexion

Digital Importer. Two librarians participating in the project also learned to catalog in MARC over the past year as part of their involvement in The Chesapeake Project. The time required to learn how to catalog and create original bibliographic records in OCLC was also cited as a challenge. Possible solutions to this challenge have been discussed by project participants, including the designation of a shared project cataloger responsible for creating more challenging original records and/or authority records, as well as permitting less-than-full level cataloging for project records, as the new CONTENTdm system provides full-text searching of PDF files, alleviating the need for comprehensive subject analysis.

Finally, The Chesapeake Project lost a valued and instrumental project leader, Bob Oakley, the Director of the Georgetown Law Library, on September 29, 2007. The Chesapeake Project would not have come to be without Bob Oakley's foresight, enthusiasm, and leadership. Project members are deeply saddened by his loss; in his absence, the strength of his vision for a nationwide legal information archive, as well as his passion for ensuring the preservation our legal heritage, continues to be an inspiration and a driving force behind the project.

C. Accomplishment of Project Mission and Vision

The mission of The Chesapeake Project is "to successfully develop and implement a pilot program to stabilize, preserve, and ensure permanent access to critical born-digital legal materials on the World Wide Web. The Chesapeake Project is working to establish the beginnings of a strong regional digital archive collection of U.S. legal materials as well as a sound set of standards, policies, and best practices that could potentially serve to guide the future realization of a nationwide preservation program."

Project participants largely felt that The Chesapeake Project has successfully worked toward achieving its mission through the project's first year (see Table 6).

To what extent do you agree that The Chesapeake Project has accomplished the following mission-stated goals during its first year?

	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Response Count
Developed and implemented a pilot program to stabilize, preserve, and ensure permanent access to critical born-digital legal materials on the World Wide Web	(0)	(0)	(0)	16.7% (1)	83.3% (5)	6
Established the beginnings of a strong regional digital archive collection of U.S. legal materials	(0)	(0)	(0)	16.7% (1)	83.3% (5)	6
Developed sound set of standards, policies, and best practices that could potentially serve to guide the future realization of a nationwide preservation program	(0)	(0)	16.7% (1)	(0)	83.3% (5)	6
	answered question					**6**

Table 6. Project participants' appraisal of the extent to which the project has met its mission-stated goals in its first year.

Most participants expressed that they felt significant progress had been made toward the project mission's call to "develop and implement a pilot program to stabilize, preserve, and ensure permanent access to critical born-digital legal materials on the World Wide Web." However, as one respondent noted, ensuring permanent access depends

on the stability and decisions made by OCLC, and as such, should be recognized as a core element of the project mission that is beyond the project's control.

Although strides have been made toward the development of a strong regional archive of legal materials, participants expressed that the project and its collections would become increasingly relevant as more materials are harvested and added to the archive in the second year. Also, while many project participants lauded policy materials developed for the project, particularly the project Collection Plan, as well as the amicable relationship and rapport built between the three participating libraries, it was acknowledged that policy issues must continue to be addressed if the project truly intends to "guide the future realization of a nationwide preservation program."

The Chesapeake Project's vision statement is as follows: "The Chesapeake Project aims to set a precedent for a national movement to prevent the widespread loss of legal information in digital formats, securing these materials for generations to come. Upon reaching the close of its two-year pilot phase in 2009, The Chesapeake Project hopes to help inspire, establish, and galvanize widespread participation in a comprehensive, collaborative, and nationwide preservation program for legal resources" (Chesapeake Project, 2007).

In assessing the extent to which The Chesapeake Project has worked toward the accomplishment of its vision, project participants recognized that the project has established a foundation for a larger project, and has made some progress in the area of setting "a precedent for a national movement to prevent the widespread loss of legal information in digital formats." However, at the close of the pilot's first year, it has not yet galvanized "widespread participation in a … nationwide preservation program for legal resources."

Most respondents stated explicitly that the goal of the pilot's second year will be to make strides in the realization of the second sentence in the project vision. Specifically, participants state that The Chesapeake Project must seek out opportunities to raise awareness of and educate others about its digital preservation work before other institutions can be inspired to participate in a larger, nationwide effort.

VI. Recommendations and Conclusions

Given the findings of this first-year evaluation, the inaugural year of The Chesapeake Project has proven to be a success. More than 2,700 digital items, representing about 1,266 titles, have been harvested from the Web archived, and roughly 8 percent of these titles have already been removed from their original locations on the Web, demonstrating the importance and effectiveness of the project's efforts. Moreover, although the project has not been marketed to users, access figures are surprisingly high, even when accounting for possible inflation of access figures by project participants. This high level of access perhaps indicates a) the effectiveness of making archived titles available via bibliographic records in OCLC and local catalogs; and b) the successful selection of high-interest and high-use materials for archiving by project participants.

Project participants themselves have also expressed satisfaction with the project's progress during the first year. Challenges posed to the project have been adequately addressed, and progress has been made toward the achievement of the project's established mission and vision statements. When asked specifically about the project's accomplishments, project participants cited the following:

- The extent and diversity of the archive, which has evolved from being empty into "a collective database of thousands of born-digital documents."
- The establishment of what may be the most comprehensive open-access collection of recently published Maryland General Assembly-mandated task force reports available online.
- The quality of the project's shared documentation.
- The enthusiasm and interest of project participants.

However, it is acknowledged that much more needs to be accomplished in the project's second year. New documentation must be created to accommodate the new CONTENTdm/dark digital archive system, and a period of training and learning to use the new system is expected to slow project progress as the second year begins. Project participants must work to maintain the momentum of the project's first year despite these new challenges.

Additionally, project participants must seek out opportunities to educate others about the project and build support for the establishment of a national legal information archive. In the coming year, The Chesapeake Project will not only launch its official Web presence, LegalInfoArchive.org, it will also work to develop new materials and policies to facilitate the addition of new participating libraries to the project. This effort will be crucial to the successful achievement of the project's mission and vision.

Prepared by:
Sarah Rhodes
Georgetown Law Library

Contributors:
Mary Jo Lazun
Katherine Baer
Maryland State Law Library

Dee Dee Dockendorf
Virginia State Law Library