THE PROBLEM

Our Digital Legal Heritage: At Risk for Permanent Loss

Written laws, records, and legal materials form the very foundation of a democratic society. Lawmakers, legal scholars, and everyday citizens alike need and are entitled to access the current and historic materials that comprise, explain, define, critique, and contextualize their laws and legal institutions. In a liberal democracy, where the rule of law requires the public to be informed, legal materials in all formats must be preserved.

Thus far, the 21st century has witnessed an unprecedented mass-scale acceptance and adoption of digital culture, which has resulted in an explosion in digital information. Consequently, important legal materials are increasingly being digitally born and disseminated online rather than published on paper.

FACT

In 2007, the amount of digital information created, captured, or replicated within a single year exceeded 281 billion gigabytes, an amount that, for the first time in history, surpassed the world's existing electronic storage capacity.

These digital items are more compactly stored, easily transportable, widely distributable, and instantly accessible than any previous information format or medium; however, their proliferation poses an extraordinary challenge for professionals in the fields of law and legal informatics who are concerned about the long-term preservation of legal information.

Of particular concern are legal materials published directly and independently to the Web, which are at an extremely high risk for permanent loss.

FACT

The longevity of digital media is uncertain. Compact disks have an estimated physical lifespan of anywhere between five to 59 years, and the descent of a given format (such as a PDF or Word file) into obsolescence, on average, can be expected to occur within five to 20 years.

Legal Information & the Web

Important legal materials are increasingly being digitally born and then distributed online rather than published on paper.

Since the mid-1990s, the number of government documents distributed in digital, as opposed to print, formats by the United States Government Printing Office (GPO) has ballooned. Federal and state agencies over the past decade have also produced a growing number of digitally born documents and reports, which have been posted directly to the Web. Court opinions are now being published online, and legal scholarship increasingly relies on digitally born sources, identified only by a Uniform Resource Locator (URL) directing to an online document.

In fact, what has been called the "disintermediation of legal scholarship" through collaborative and open-access Web-based publishing is having a noticeable impact on the practice and study of law in the United States. Articles and commentary posted on legal Web logs (blogs or "blawgs") have been cited in many prestigious law reviews as well as in cases argued before state and federal courts, including the United States Supreme Court.

FACT

In 2005, Law Professor Douglas Berman's *Sentencing Law and Policy* blog had the distinction of being the first blog cited by the United States Supreme Court. The citation appeared in a dissenting opinion issued in the landmark case, *United States v. Booker*.

The transient quality of legal information published directly to the free Web (as opposed to within subscription databases), often by government and independent entities, is troubling.

Documents, reports, and other legal information published online 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.

FACT

Studies have shown the average lifespan of a Web page to be between 44 and 75 days.

THE SOLUTION

The Chesapeake Project Legal Information Archive

The Chesapeake Project's Legal Information Archive is a collaborative, two-year pilot digital preservation program established to preserve and ensure permanent access to vital legal information currently published in digital formats on the World Wide Web.

The project was implemented in early 2007 under the auspices of the Legal Information Preservation Alliance (LIPA), an independent organization supported by the American Association of Law Libraries. The Chesapeake Project is being carried out by three LIPA-member libraries:

- The Georgetown Law Library
- The Maryland State Law Library
- The Virginia State Law Library

Discovery & Access

Discovery of and access to The Chesapeake Project's archived collections is made available through:

- Participating libraries' local catalogs
- The open-access WorldCat.org system
- Subscription OCLC FirstSearch and WorldCat databases
- The Chesapeake Project's new CONTENTdm system

As a digital item is harvested from the Web and archived, it is assigned a unique URL hyperlinked to the archived access copy in the OCLC system. This URL is added to bibliographic catalog records, providing direct access to archived objects. If and when an object's original URL becomes inactive, the URL for the archived access copy will continue to provide access to the title. Any user with an Internet connection can discover these records through traditional catalog searching methods, using a library's OPAC or an OCLC database, and is provided with open access to archived resources.

Digital Archiving Strategies &Tools

The libraries participating in The Chesapeake Project harvest content from the Internet and preserve it, along with the appropriate preservation metadata, within a shared digital repository. While some Web-harvesting projects focus on the capture and preservation of entire Web sites, The Chesapeake Project focuses upon the capture and preservation of discrete online publications.

Original System: OCLC Digital Archive

The Chesapeake Project initially utilized the OCLC Digital Archive, operated and administered by the nonprofit Online Computer Library Center (OCLC). The OCLC Digital Archive adheres to the ISO reference model for an Open Archival Information System (OAIS), the standard conceptual framework for the permanent preservation of digital information. Archived files remain uncorrupted and renderable in their original formats. OCLC provides secure onsite storage of archived items at OCLC facilities, maintaining multiple copies of backup data and disaster tapes stored at an offsite facility, and a regular schedule of virus-checking, file format verification, and fixity-checking using checksum algorithms. Technical metadata are automatically generated, including the file format type, verified using the JSTOR/Harvard Object Validation Environment (JHOVE). Project participants play a curatorial role in the creation of metadata records, manually entering descriptive and administrative metadata into the preservation records.

Current System: CONTENTdm + (Dark) Digital Archive

In April 2008, OCLC began transitioning The Chesapeake Project's archived collections and metadata from the original OCLC Digital Archive to a more sophisticated, two-tiered digital-preservation and access system. Using the new system, two digital objects are created from the original item harvested from the Web: a master file and an access copy. The master file is stored in a dark digital archive, very similar to the previous OCLC Digital Archive, except that it is completely inaccessible to users. The derived access copy is imported into CONTENTdm, a customized storage and retrieval system, which makes archived collections accessible to users via a searchable Web interface.

FIRST-YEAR SUCCESS

First-Year Pilot Project Evaluation Results

In April 2008, The Chesapeake Project conducted and published a formal evaluation of the project through its first year, spanning from February 27, 2007, to February 29, 2008. Evaluation findings included the following:

» More than 8% of archived titles already missing from original Web locations

More than 8% of the titles archived 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.

» More than 2,700 items harvested from the Web & archived

Since the project began, the 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. 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.

» Archived items accessed more than 5,300 times during first year

Despite the fact that the project was not publicized during its first year, ***archived items were accessed a total of 5,317 times***. Although project participants, during quality control link-checks, accessed their own items 2,267 times, public users accessed project materials through open-access links ***a surprisingly high 2,528 times.*** Other libraries and institutions, excluding project participants, accounted for ***522 instances of access***, occurring during the course of research, reference activities, and adding archived URLs to their own local catalogs.

» A successful, flexible collaborative project management model established

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 5 to nearly 70. 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.

» Project participants are enthusiastic about continuing past the pilot phase

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.

Looking Forward

It is important to remember that The Chesapeake Project is a two-year pilot, and it ultimately aspires to evolve into a much larger digital archive for legal materials, shared by law libraries throughout the United States. With the organization-wide support of the Legal Information Preservation Alliance and the American Association of Law Libraries, this vision is indeed within reach. Beyond the borders of the United States, The Chesapeake Project aims to inform the preservation initiatives of other organized groups of libraries, who may learn through its experiences, and to raise global awareness of the vulnerability of digitally born legal materials published on the Web.

Presented by:
Sarah Rhodes
Digital Preservation Librarian
Georgetown University Law Library
(202) 662-4065
sjr36@law.georgetown.edu

FOR MORE INFORMATION

The Chesapeake Project Legal Information Archive (beta CONTENTdm site)
» legalinfoarchive.org

The Chesapeake Project First-Year Pilot Project Evaluation
» legalinfoarchive.org/cdm4/policies/LEGAL_FirstYearProjectEvaluation.pdf

The Chesapeake Project Collection Plan
» legalinfoarchive.org/cdm4/policies/LEGAL_CollectionPlan0907.pdf

Legal Information Preservation Alliance
» www.aallnet.org/committee/lipa

CONTENTdm Digital Collection Management Software by OCLC
» www.contentdm.com

OCLC Digital Archive
» www.oclc.org/digitalarchive

WorldCat.org (open access)
» www.worldcat.org