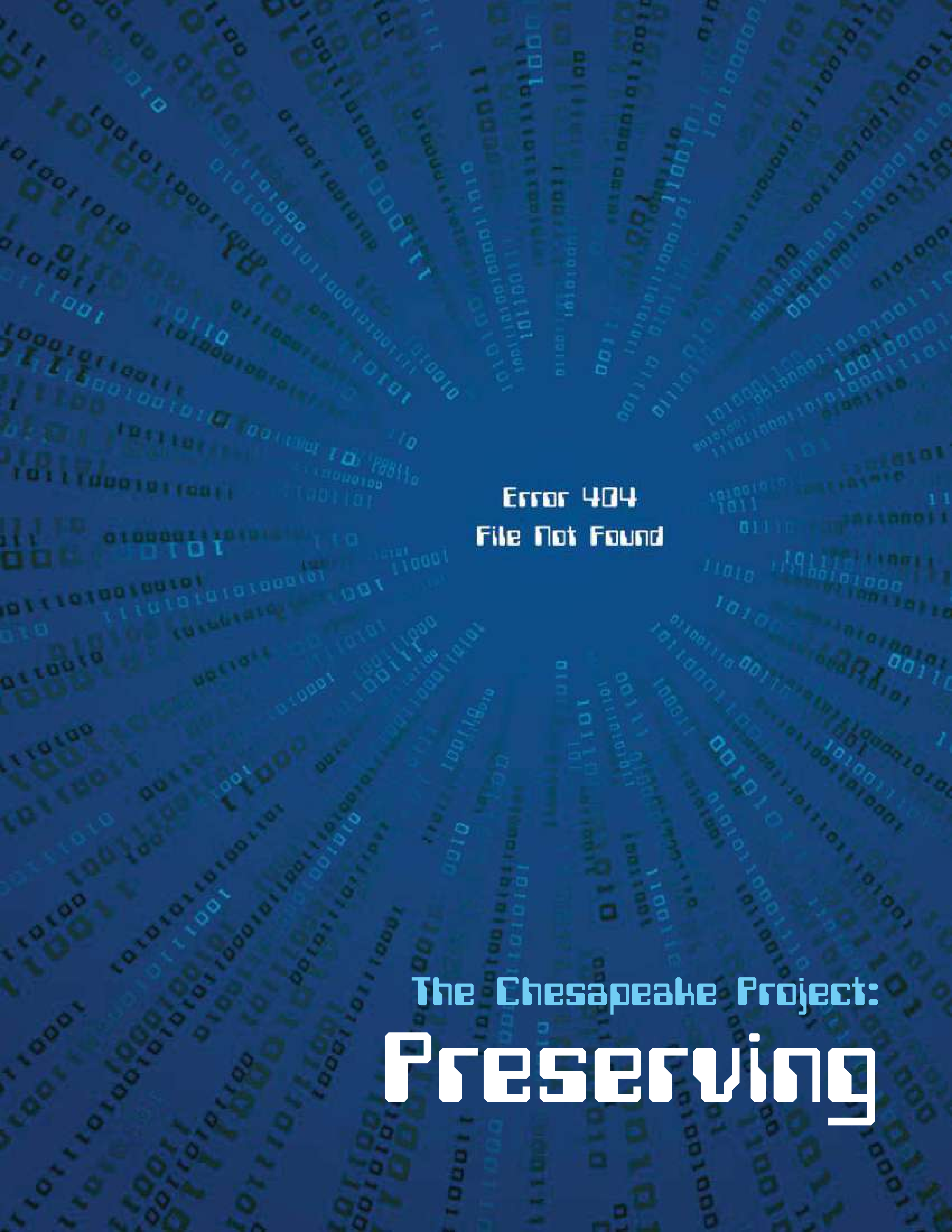
Georgetown University Law Center
Scholarship @ GEORGETOWN LAW

2009

The Chesapeake Project: Preserving the Digital Future

Anne Cassidy
Georgetown University Law Center, acc52@law.georgetown.edu

Georgetown Law Alumni Magazine, Fall/Winter 2009



Error 404
File Not Found

The Chesapeake Project: Preserving

the Digital Future

By Anne Cassidy

The New York Police Department report "Radicalization in the West: The Homegrown Threat" made a splash when it came out in August 2007. Discussed on blogs and in the *New York Times* and other publications, the report examined terrorism cases in New York City and elsewhere in an attempt to understand what drives "unremarkable" people to become terrorists. Like so much born-digital information, however, the report is no longer available on its original URL — and if you find it elsewhere, its authenticity may be compromised. But thanks to the Chesapeake Project, a digital preservation effort spearheaded by the Georgetown Law Library in conjunction with the state law libraries of Maryland and Virginia, you can access the original report and thousands more like it.

The digital universe is growing at a mind-boggling rate. If you were to print out its estimated 500 billion gigabytes of information into books, the volumes would stretch to Pluto and back 10 times. Preserving and storing digital information has become a major challenge for companies, governments and universities around the world.

If you were to print out the estimated 500 billion gigabytes of information [available on the Internet] into books, the volumes would stretch to Pluto and back 10 times.

Sometimes it seems like a losing battle. A study in the *Journal of Information Science* found that three-quarters of the more than 700 Web site URLs it followed between 1997 and 2004 disappeared from their original locations during that time. Web sites vanish during routine maintenance, because of a change in management or administration or when Web masters update URLs. In some cases, this may be only a minor annoyance, but when Web-born information is cited in legal journal articles or used in court cases, it's especially important to guarantee longevity.

"The average lifespan of a Web site is 44 to 75 days," says Georgetown Law's digital preservation librarian Sarah Rhodes, who manages the Chesapeake Project. But a funny — and telling — thing happened when she tried to source that fact. "I first saw this statistic on the Library of Congress's digital preservation page. And then I saw it again on an Internet Archive page," Rhodes says. "I looked and looked for the source of this statistic and found that it was a study that was only published online. The original version is no longer available."

It's to avoid such occurrences that the Chesapeake Project was born. While it's by no means the only such project of its type, it's certainly one of the most comprehensive. More than 4,300 digital items were harvested from the Web and preserved in the project's archive during its first two years. "I wouldn't say we're ahead of others [in harvesting digital information], but we're definitely in the forefront," says Janice Anderson, Georgetown Law's associate librarian for collection services.

Anderson has been involved in the Chesapeake Project from the beginning. In 2003, the Law Library, under the directorship of the late Bob Oakley, gathered publishers, library directors, archivists and others for a conference called "Preserving Legal Information for the 21st Century: Toward a National Agenda."

"We were concerned that enough wasn't being done by law libraries to preserve digital legal information," says Anderson, who was at the meeting. From this gathering grew the Legal Information Preservation Alliance, dedicated to providing the organizational framework and professional commitment to bring about a "national consistency" in the preservation of print and electronic legal information. "Nothing less than transmission of the permanent, accurate record of legal knowledge to future generations is at stake," reads LIPA's mission statement.

While the Chesapeake Project began under the auspices of LIPA, it was Georgetown Law that spurred it on, recruiting Maryland and Virginia to join the effort — because partnership is crucial to meeting such lofty preservation goals — and hiring Rhodes, who has devoted much of her time to the project for its first two years. The name "Chesapeake" comes from the D.C., Maryland and Virginia partnership.

How Harvesting Happens

It is no small irony that in seeking to name various technological tasks we adapt the language of the past, even the long-ago past of our agrarian ancestors. The highly skilled librarians who select and safeguard Web content, therefore, are said to "harvest" it. But before they harvest, they must comb the Web to learn what is endangered. They do so by keeping abreast of news and legislative services and by monitoring the availability of state documents. They may subscribe to a resource such as WatchThatPage.com, which notifies users whenever pre-selected Web pages are changed. Often in looking for one document librarians find another, and another and another. In some cases — increasingly more often these days — they hear directly from a publisher that a report's hard copy days are over. And this, given the short life span of Web sites, puts the publication on a de-facto warning list.

"I may get an e-mail saying, 'We're not printing this anymore — we're publishing it on the Web," says Mary Jo Lazun of the Maryland State Law Library. "If we don't harvest the Web version [of that document] soon we know that within six months or a year or two it's going to disappear or move."

Due to the unique nature of legal documents, 95 percent of the material harvested is in PDF format. But the type of document depends in part upon the entity doing the harvesting. The Virginia State Library primarily harvests circuit court reports and annual state-of-the-judiciary reports. Katherine Baer of the Maryland State Law Library says she preserves Web-born documents produced by various agencies, task forces and governor's commissions — most anything produced by state entities. "We go pretty broad, anything that may be related to the law, but we also get annual reports for state agencies, various statistical reports, that sort of thing."

The Georgetown Law Library leads in the harvesting effort. "We're pulling documents from all over the place — the American Bar Association, the Urban Institute, the Department of Homeland Security, the U.S. Department of Justice and the Conference of Mayors," Rhodes said during an interview, as she perused the Web site of the project's digital repository (www.legalinfoarchive.org) from her office in Georgetown Law's Williams Library. Of special interest are journalism, copyright law, public health law, environmental law and policy, conflict resolution and problem solving, human rights and the Supreme Court — all areas of Law Center expertise.

Librarians must certify that the material they're harvesting is authentic, which the Law Library does via a periodic review of files. One way to do this is with a program that adds up all the 1s and 0s that make up a digital file, creating a number called a checksum. "If there's a change in that number [from one measurement to the next] then that file is considered corrupt," Rhodes says. The librarian also creates descriptive metadata for each file — including what format to use to retrieve the document, which law library harvested it, whether there are any permission or copyright restrictions and other information.

A study in the *Journal of Information Science* found that three-quarters of the more than 700 Web site URLs it followed between 1997 and 2004 disappeared from their original locations during that time.

Fully Searchable

When the Chesapeake Project began, no one could have imagined the hurdles it would face — beginning with Oakley's death, at the age of 61, in September 2007. "Bob was in a lot of ways the vision behind the whole project, and losing his leadership was really something. But we've worked together well since then," Rhodes says.

When the project began, Rhodes and others had to decide how the preservation piece would be handled. Rather than using their own homemade digital repository with open-source software, they decided to use the Online Computer Library Center, the nonprofit computer library service and research organization that also produces WorldCat, the worldwide library catalog.

"One year into the project we were notified by OCLC that they were migrating our materials to a new, more sophisticated archive," Rhodes explains. So only months after learning one system of cataloging and coding, Rhodes had to learn another, the Contentdm digital collection management software. But this was fortuitous, she says. Whereas with the previous system, users tapped directly into the archived copy of a file, the new system provides users with an access copy, while the original remains safely archived. "It's a two-tiered system," Rhodes explains, "with a dark archive on the back end, which the user cannot access, and a content management system on the front end."

During the two-year pilot phase of the Chesapeake Project, almost 16 percent of the information harvested from sites with a .state or .us domain vanished from the original Web location — and 26 percent of the items with an .edu address disappeared.

Another advantage to this system is that it's fully searchable by Google and other search engines, as well as via library catalogs, which is the only way it was available before. This made user figures soar so high (from 6,612 hits to 177,152) that at first Rhodes thought there was some mistake. But there wasn't. Rhodes knew then that there was a tremendous audience for the material she was preserving.

Petabytes and Link Rot

During its first two years, the Chesapeake Project has grappled with issues familiar to digital preservationists everywhere — starting with what to preserve, which Rhodes likens to fly fishing. "We're identifying very targeted things that we're preserving, as opposed to throwing a whole net in and capturing a gigantic Web site that has all sorts of different files." But such pinpointed preservation entails countless decisions and much collaboration. One important philosophy behind the Chesapeake Project is that "one library couldn't do this alone," Rhodes says.

And then there's the storage issue — and the world of terabytes (1000 gigabytes), petabytes (1000 terabytes) and beyond. Because an outside vendor is hosting the Chesapeake Project, it has been spared some of these concerns, but every preservationist is keenly aware of the space it takes to store vast amounts of digital material. At a conference hosted by the Library of Congress two years ago, Ken Thibodeau, director of the Electronic Records Archives system for the National Archives and Records Administration, said ERA would require 250 petabytes of storage space, which was more than could be mustered at the time.

State governments, the Library of Congress Web Archives, the Government Printing Office's Federal Digital System (FDsys) — all these organizations are grappling with the best ways to store vast amounts of digital data. As for protecting the data once it's stored, OCLC keeps files at two different geographic locations and at any given time there are multiple copies of each file.

Obsolescence is another concern. Although the PDF format seems relatively stable and flexible now, "who knows whether it will exist in 20 years," Lazun says. There are two ways of looking at this issue, Rhodes explains. One is to transfer files to new formats so they can be accessed (much as you would migrate old home videos from VHS to DVD). Another school of thought says that the authenticity of the original is lost in that process and that it's purer to use emulator software that allows the old to function within the new. (It's emulators that allow Pac-Man fans to play their favorite game in its "original" format.)

One of the most challenging aspects of digital deliquescence is the colorfully named condition known as link rot, which describes what happens when you click on a link and get a "file not found" message. During the two-year pilot phase of the Chesapeake Project, almost 16 percent of the information harvested from sites with a .state or .us domain vanished from the original Web locations — and 26 percent of the items with an .edu address disappeared.

Rhodes was surprised. "You would expect a .com or a .org to be less stable than a state government URL," she says, "but we found that a lot of these state government publications were also disappearing." The Chesapeake Project's first-year evaluation found that 8.3 percent of the titles archived between February 27, 2007, and February 29, 2008, were inactive by their one-year mark. And by the second year of the project, URL inactivity had increased to 14.3 percent. In one year's time, URL inactivity increased by 73 percent. By March 2009, one in seven of these URLs had become inactive.

Still, the librarians of the Chesapeake Project know they are helping stem the tide of digital disappearance. "One of the successes of the Chesapeake Project is that there are so many things that have been harvested that are no longer on the Web," says Dee Dee Dockendorf of the Virginia State Law Library. "It's nice to know we're doing something useful."

Needle in the Haystack

Now that the Chesapeake Project just ended a successful two-year pilot stage, it is set to become a model for user-friendly digital preservation. Not only are the archived documents important for libraries and researchers, but "a lot of students are using them, too," Rhodes says. The project is reaching even further than it might because catalogers are taking the records and downloading them into their system, which means they're putting digital archive links directly into their catalog. This is "really what we want them to do," Rhodes adds.

The goal, she continues, is to evolve into a regional digital archive. "We may find that a single project like ours can become a national project. Or we may also find regional pockets that spring up all over the country. We like to think that our project will provide a foundation for other projects to build on in the coming years."

It makes sense that the preservation of legal information might serve as a catalyst for other forms of digital protection. "I'm not sure if law librarians are leading the way here," Dockendorf says. "But I do think law librarians are a little more concerned [about preservation] because a lot of times it's that needle in the haystack that people look for that can really affect a case or argument."

Even though Rhodes has been mired in the everydayness of the Chesapeake Project for the last two years, she knows that what the project is really about is the future. "The real value of this is how we might look back at it in 20 years," she says. "When I look at the big picture it feels like we're doing something for the future. I think it's a noble project."

> "When I look at the big picture it feels like we're doing something for the future. I think it's a noble project."

Preserving the Blogosphere

The research value, reliability and preservation of blogs were the topics discussed at "The Future of Today's Legal Scholarship," a symposium held Saturday, July 25, at Georgetown Law. Librarians, law professors, and electronic preservationists from the Library of Congress and elsewhere contributed to the conversation. Librarian Margaret Schilt of the University of Chicago noted that blogs serve important functions in terms of public policy discussions and can be a testing ground for later scholarship. How blogs are going to be kept available over time — especially when they're being cited in cases and scholarly articles — is another question. The symposium was held in honor of Professor Bob Oakley, director of the Georgetown Law Library from 1982 until his death in 2007. "Those of you who knew Bob recall his deep interest in promoting access to legal information and the responsibility that librarians have to preserve and authenticate that information," said librarian Peggy Fry.

—Ann W. Parks