

## The Need for Digital Archiving Standards

By Michael Looney

Campus tour guides at Yale University are known to tell a story about Yale's Beinecke Rare Book and Manuscript Library, one of the world's great document repositories and home to a copy of the Gutenberg Bible, the first Western book printed from movable type. Standing outside the Beinecke, the guides describe a remarkable mechanism that, should the terrible need arise, would cause its glass-encased central tower and its 780,000 volumes to withdraw deep underground, tucked away from any possible threat of destruction.

It isn't true, but when visitors see the illuminated pages of the Gutenberg Bible or peruse the papers of Samuel Clemens, the idea of a subterranean vault sounds prudent.

Librarians and archivists at colleges across the country couldn't agree more. Storing important works so future scholars have access is a vital part of what many university libraries and museums do. But as print-based publishing is outpaced by an onslaught of digital material, traditional archival methods are facing challenges.

The job of digitally storing and sharing that content is increasingly complicated. The Web, as just one example, is the largest living document ever created. At four billion public pages (and another 550 billion pages accessible via the "deep Web"), it is 55 times larger than the entire contents of the Library of Congress. Only 10 years old, the Web already is a fundamental resource for students and faculty, who find they are moving from a print-based world to one saturated with ever-changing digital content. The Web adds seven million new pages every day, but on average those pages disappear in 44 days (Lyman 2002).

Fortunately, universities and research libraries are committed to keeping up. They increasingly are incorporating digital information sources into their collections and curricula, while making digital records of physical archives. Yet they are also finding that the old, universally accepted ways to catalog and access information often no longer apply. Archives aren't purely physical places for the archivist, librarian or faculty member conducting research or teaching.

### The Standards Conundrum

We can all be thankful that, through the ages, we've collectively decided what form a book should take. In Western civilizations, we know that we can find the title on a book's spine; we safely assume the contents will be printed on sheets of paper, ordered from left to right; we know how to use the college library with varying degrees of success.

Now, imagine if these little details weren't decided at all, leaving it to individual publishers to decide what a book really is. Or, if each college librarian created his or her own cataloging method, forcing us to completely re-learn the process of locating information from one library to the next.

It would, of course, be chaos, which is precisely what standards are designed to prevent. Nowhere is this more relevant than in the creation of digital archives, whose future is dependent upon standards—community standards that define common procedures, and technology standards that uniformly enable digital storage and retrieval. Pinpointing best practices and technologies for digital archiving is a core initiative of RLG, a consortium of 160 universities, national libraries, and other institutions. "Without standards, there is really no hope for digital archives to be usable many years from now," notes Merrilee Proffitt, program officer at RLG.

The discussion of standards can get overwhelming in no time. For instance, consider the quandary of the digital archivist, who must determine the right processes and technologies to create digital records of printed works, films, audio tapes, or images. Those creating archiving structures also must define easy ways for students, with little or no training, to quickly peruse their options and locate exactly what they need—preferably without first having to open a single digital file.

These issues tie directly into questions about technology. What format that exists today will sustain digital records for 20, 50, even 100 years? Content must be easy to capture, and it must be viewable using tools that are readily available. What's more, a file created in 2003 must be viewable in 2099.



### SMU Dives into Databases

Law students at SMU's Dedman School of Law are undergoing their own digital revolution, in the wake of a similar transformation among leading professional law firms. Adjunct Professor Steve Kardell, who is also a corporate governance attorney in Dallas, supplements classroom materials with content from the online LexisNexis database and Web pages captured via the Web capture feature in Acrobat 5. "It's amazing for students to see how much content is out there in addition to standard classroom material," says Kardell.

Acrobat also simplifies the process of keeping archived materials up to date. "These are living documents, so we need a way to keep them

current," says Kardell, who also is evaluating requirements for a permanent SMU digital law archive. "With Acrobat, authorized editors just go into the portal from time to time and update the material. You can highlight and annotate material in ways that people don't realize."

In the transient world of computer technology, that's a tall order. For instance, there's ASCII, the only electronic document format recognized by the National Archives and Records Administration. ASCII does a fine job of recording and displaying text—so long as it isn't Asian text, which requires more than ASCII's set of 128 characters. In fact, ASCII's usefulness stops well short of many materials headed for digital archives. It can't, for instance, accurately render a Web page featuring photos and reports from the World Trade Center attacks on Sept. 11, 2001.

HTML, the language used to display Web text, goes one better by formatting text into layouts and identifying areas for photos or links. But not all HTML is created in precisely the same way—an HTML page may appear differently on your Web browser than on mine. When researchers and students need to study an item exactly as it appeared, the limitations of HTML become apparent.

In settling on the right technologies, archivists must match expected longevity with visual acuity. Peter Ullmann, a key participant in standards efforts at Adobe, puts it simply: "What's going to give you the sense that you're looking at something real?"

Though lacking a universally accepted standard, many schools already are hard at work building digital archives for curricular and administrative use. Typically, they establish their own processes and select the technologies that best meet their needs. Many of these systems incorporate the MARC (MACHine Readable Cataloging) system developed in the 1960s by the Library of Congress. The MARC system provides electronic access to bibliographical information for a library's inventory, and may serve as a model for much of the "metadata"—or contextual information—that will tag digital archives of tomorrow.

But that's tomorrow. Many universities already have successfully created distinctly different digital archives with Adobe Acrobat software, a low-cost authoring tool that easily generates any document in Portable Document Format (PDF) (see "Electronic Archives at Whitman College"). PDF is a broadly accepted, open specification for final-format documents that can be viewed using freely available Acrobat Reader software. PDF retains the format of the original document or Web page, so elements like pagination, photographs, and hyperlinks remain true to the original.

### Electronic Archives at Whitman College

Professors at this small, liberal arts school in southeast Washington often require their students to use materials that are not available in the public stacks at the 385,000-volume Penrose Library. "A professor might have material he's gleaned over the years that he wants his students to have access to," says Michael Quiner, director of administrative technology at Whitman. "The old approach is to make photocopies, and then students go to the Reserve Desk to read them." For greater flexibility, Whitman administrators devised eReserve, a growing digital archive of reserve materials viewable online from students' dorm rooms.

In two years, the eReserve program has archived some 400 articles as PDF files that authorized students can read using a standard Acrobat Reader. In fact, Quiner says, several professors have started their own efforts to archive class material in PDF so students can conduct research online.

To College Librarian Henry M. Yapple, the benefits extend beyond allowing a student to read documents at all hours: "It's a matter of preservation," Yapple says. "eReserve allows a lot of people to read the documents without handling the primary material." Yapple's argument rings especially true with documents like Yale's Gutenberg Bible, whose rare pages could be viewed digitally by any religious studies student without ever touching the book.

### Where We're Headed

While organizations like RLG work to define digital archiving standards, certain technologies are likely to find themselves at the forefront of the debate. Their prominence suggests that they will play at least some role as standards evolve.

One of these technologies is XML (eXtensible Markup Language), which is becoming vastly popular for many applications. XML allows information to quickly come together from various locations to form a Web document that can be easily read, and features an advanced approach to tagging content so that its components appear in their logical order once they reach their destination. XML appears to be an excellent candidate for supplying the technical backbone of a digital indexing system. "XML schema language can provide the universal structure that allows any school to look at technical metadata," explains RLG's Proffitt.

Maintaining accurate page format of the paper document, however, is not among XML's many strengths. This doesn't matter if a student is reading the text of Dr. Martin Luther King Jr.'s historic "I Have a Dream" speech. But XML is at a disadvantage when a student must view an original document.

While XML excels at transporting information, PDF excels at displaying visually rich information. PDF preserves the pagination integrity of original documents, even when they are viewed on PDAs or next-generation wireless phones. Digital archiving is a marriage of data and documents. The two must live together, and for a very long time.

Adobe's recent development around PDF recognizes this. Acrobat 5.0 exports XML along with PDF, resulting in an XML-tagged

document that retains its pagination no matter how it is reviewed. This would allow a journalism student to look up a story from yesterday's Los Angeles Times on her handheld. She knows the story appeared on page 16 in the print version. If the page is stored in PDF and tagged with XML, that's exactly where she'll find it on her PDA.

As new display devices become popular, these capabilities will be necessary to find and view records that were archived years before. Wharton's Kendall Whitehouse already has proven this is possible with PDF, viewing a document archived in 1995 on three different platforms: a desktop computer, a Palm OS handheld, and a Compaq iPAQ Pocket PC. When Whitehouse archived the document, those two handheld devices did not even exist (see "Wharton Students Straight to PDF").

### Wharton Students Straight to PDF

In 1993, Wharton administrators began digitally preserving all school publications, catalogs, course materials, and faculty research papers. Yet the school's efforts don't stop at archiving. Some faculty members also have their students electronically submit their assignments as PDF files, which are then annotated with comments and corrections and re-posted for the student to view.

According to Kendall Whitehouse, Wharton's director of advanced technology, a successful archiving format must faithfully represent the original work without requiring complicated technical back flips. "A lot of formats depend upon the material being created in a certain way," says Whitehouse. "But Acrobat and PDF are completely agnostic. You can view the files on Windows, Mac, Unix, and handheld devices. None of these require manual altering or modification."

Archives must be easily viewable for years to come, a constraint that automatically narrows the technology field. "Our original PDF documents from 1993 have actually improved with age because the Acrobat Reader has evolved," Whitehouse says. "It's hard to list other formats that are both backward and forward compatible."

An indexing feature called e-Binding also allows educators to combine multiple content in various formats—images from a photo essay, maps from an atlas, spreadsheets, and text documents—into a single PDF file. And several legacy documents can be batch-processed to create a multi-layered work that can be searched for key words or phrases. For faculty, this offers the chance to create an online "course pack" for students that is easily indexed, searchable, and updated over time.

While PDF itself has become a de facto industry standard, two industry organizations are jointly working to establish an official archiving standard based on PDF technology, called PDF/A (see "Making the Case for PDF/A"). The groups are working to see PDF/A recognized by the International Standards Organization as a global standard for document archiving.

For now, digital archivists seem focused on tackling the issue of electronically documenting old and rare printed works, or capturing a Web page before it changes only a few hours later. And as immensely useful as digital archiving standards undoubtedly will be, educators point out that a ubiquitous system for higher education won't likely replace fixtures like research librarians anytime soon. "Tracking down information correctly is a tricky business, and it requires a skilled professional to do it," says Whitman College Librarian Henry Yaple. "That's what librarians do." With solid digital archiving standards, that job may become considerably easier.

### Making the Case for PDF/A

A recent study estimates that the world's total production of information amounts to about 250MB—some 100,000 pages—for each man, woman, and child on earth. Printed documents comprise only .003 percent of the total (Lyman and Varian 2000).

99.997 percent of all information is digital—and it's growing fast. Some futurists anticipate that someday the world's knowledge will double every 900 days. The Census Bureau, for example, has accumulated 600 million pages of information from the 2000 Census that it will be transferring to the National Archives and Records Administration (NARA)—equaling 10TB of data. That's more than five times the amount of data that NARA has captured and fully processed in its entire 30-year history.

Yet because of their historical value, billions of documents need to be managed, preserved, and made accessible for future generations. This daunting task requires a solution that recognizes the wide range of information systems, technologies, and formats in which records are generated.

To a growing number of industry groups and users, one solution is PDF—a broadly accepted standard for the delivery of final-format documents. More than 20 million PDF documents are publicly available on the Internet, and almost half a billion copies of the free Acrobat Reader have been downloaded. PDF retains the content, look, and feel of the document exactly as it was created, ensuring document integrity and security, while also allowing documents to be searched. In fact, some countries already have accepted PDF as an archive standard. However, PDF has evolved to provide a number of functions that, while beneficial to users who share documents, are not ideal for long-term archiving: password-based security of documents, optional (rather than required) embedding of specific fonts, the ability to embed multimedia in other formats, and the ability to launch other applications from within PDF.

Consequently, a subset of PDF—PDF/A, with the "A" standing for archive—is being developed for archiving and preserving digital documents. PDF/A—a joint initiative by the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES) and the Association for Information and Image Management, International (AIIM)—will address the growing need to electronically archive documents to ensure preservation of their contents over an extended period of time. PDF/A will also ensure that those documents can be retrieved and rendered

with a consistent and predictable result far into the future.

PDF/A proponents (a working group comprised of industry, government, and academic institutions working with AIIM and NPES) are aiming to have PDF/A officially recognized by the International Standards Organization (of ISO 9000 fame) within approximately 18 months. Their efforts are directed at solving a serious and increasingly urgent problem. The lack of a recognized and accepted electronic standard for records preservation—particularly as new generations of hardware and software have made previous digital technology obsolete—has led to the loss of significant amounts of valuable information over the past several decades. Military files from the Vietnam War, records from the Viking Mars Mission, Census Bureau data and land use records have been lost due to the inability to read data formats and the deterioration of magnetic tapes used to store that data.

The list of organizations that mandate or use PDF as a de facto standard is growing to include the U.S. Courts, the National Science Foundation for grant submission, and the Food and Drug Administration for drug submissions. A common PDF/A standard will give librarians and educators the confidence that their records could be readily accessed far into the future.

## Resources

Lyman, Peter, "Archiving the World Wide Web." *LOOP: AIGA Journal of Interaction Design Education*, December 2002, Number 6. Retrieved from <http://loop.aiga.org/content.cfm?ContentID=100> on Jan. 22, 2003.

Lyman, Peter and Varian, Hal R., "How Much Information," 2000. Retrieved from [www.sims.berkeley.edu/how-much-info](http://www.sims.berkeley.edu/how-much-info) on Jan. 26, 2003.

Michael Looney is a senior director at Adobe Systems Inc.

This article originally appeared in the [3/1/2003](#) Issue of Syllabus

**CAMPUS TECHNOLOGY** : [Magazine](#) | [Conferences](#) | [News](#) | [Community](#) | [About Us](#) | [Advertising Information](#) | [Home](#)

[Application Development Trends](#) | [Campus Technology](#) | [CertCities.com](#) | [The Data Warehousing Institute](#) | [E-Gov](#) | [ENT News](#)  
[Enterprise Systems](#) | [Federal Computer Week](#) | [IT Compliance Institute](#) | [JavaSPEKTRUM](#) | [TechMentor Conferences](#)  
[MCPmag.com](#) | [OBJEKTSpektrum](#) | [Recharger](#) | [Redmond magazine](#) | [SIGS-DATACOM](#) | [TCPmag.com](#)

Copyright 1998-2004 101communications. See our [Privacy Policy](#)

