WHITE PAPER

PDF as a Standard for Archiving

The Portable Document Format (PDF) as a solution for creating archives for both paper and electronic documents

Benefits and Requirements for Electronic Archives

Today's Archiving Challenges

If you suddenly couldn't access the building where your vital records were kept, what would you do? In today's business and political climate, corporations and governments around the world are beginning to pay close attention to their processes and the resulting record archives that they are—or are not—keeping. In a paper-based world, traditional archiving has meant storage of paper, but what happens as more and more records are created electronically? How do you preserve both paper-based and electronic records in a consistent format? How do you eliminate the need for paper records? How do you preserve the exact look and feel of a document today or 30 years from today? How do you provide consistency in the integrity of your archives?

The introduction of personal computers into business has drastically changed the archiving environment. Prior to the 1990's, most offices still had typing pools and word processing groups and kept records on paper in centralized files. But once computers became the norm for the majority of workers, the usefulness of the centralized file room disappeared. It became everyone's responsibility to create, file, and maintain his or her own documents. As a result, corporations and governments have lost control over these records.

Benefits of Electronic Archives

The advantages of establishing and continually building an electronic archive for an organization are numerous. Electronic archives unlock information that was previously difficult to access in paper form, enable more effective sharing of information, and contribute to knowledge flows. No longer are archives the domain of those few who truly understand the filing system. With electronic archives, information can be made available to anyone in the organization by granting access privileges.

Additionally, electronic archives can contribute to extensive cost savings within an organization. The cost associated with maintaining paper-based archives can be great, and electronic archives can help to significantly reduce this cost. For instance, a well-known study from PricewaterhouseCoopers LLP found that for every 12 filing cabinets in an organization one additional employee is required to maintain them. Further, while professionals spend only 5% to 15% of their time reading information, they spend up to 50% of their time looking for it.

TABLE OF CONTENTS

Adobe

Establishing Requirements for Adequacy of Records

Electronic archives can provide reliable evidence of past actions and decisions, but to do so they must be managed so as to retain the integrity and authenticity of the records. Achieving this goal requires paying attention to the records management program and expanding currently held definitions of records to encompass not only paper but electronic records and other media as well. Establishing and maintaining an electronic archive requires policy decisions, procedures, and organization-wide planning along with a commitment to follow the organizational standards.

Preserving the content, context, and structure of records is not a new concern for records management. Luciana Duranti of the University of British Columbia (UBC) employed the science of diplomatics as the theoretical foundation of electronic records research, using the rules of diplomatics to establish the reliability and authenticity of electronic records. UBC's research concluded that organizations need to use the same records policies and procedures regardless of whether the record is created on paper, kept electronically, or converted to microfiche. By treating all records in the same manner, the authenticity and integrity of the records are enhanced.

The requirements for the adequacy of records are determined by each organization's internal business and legal needs, as well as external regulations or requirements. Thus, the requirements for each organization will be different. A thorough risk analysis must be performed with the full participation of the organization's legal department to determine the technological approach that is right for that organization. The assessment team should include:

- Auditors and lawyers: Knowledge of the organization's business structure, procedures, and laws and policies that apply to the organization's records
- Records managers and archivists: Knowledge of who accesses the records, why the records are accessed, and how long the records need to remain accessible
- Record creators and users: Knowledge of the records' business purpose and operational value

Requirements in the Context of PDF Files

Records managers and archivists believe that records must be authentic, reliable, complete, unaltered, and usable and that the electronic systems that support the records must be able to protect their integrity over time. But what does this mean, especially as it relates to PDF files?

- Authentic: It must be possible to prove that a record is what it purports to be, that it has been created or sent by the person who claims to have created or sent it, and that it was sent at the time alleged. This can be accomplished by use of metadata, which is data about the data. In the case of PDF files, metadata can be programmatically embedded inside of the file, thereby ensuring that it is what it purports to be. The creation, receipt, and transmission of records need to be controlled to ensure that record creators are authorized and identified. While this is usually a function of the overall electronic records management system, there are certain features of PDF files, such as security settings, that support the establishment of authenticity. Electronic signatures are an additional level of authenticity that can be applied to PDF files.

In electronic transactions, a PDF file can combine data on who used the system, when they used it, what they did while using it, and the results of the transaction. A savvy programmer can use PDF files to capture and preserve as many elements of the electronic transaction as possible, specifically, the complete "visual presentation" of the transaction to the user. Sometimes this visual presentation is shown back to the user for confirmation. This can significantly improve an organization's ability to meet the tests of admissibility, since the record of the transactionTw (to 31.5424(cyd)002(s)-5.7of ad-15.7(t)-0.of t)-35san3i 4 Tw (w)-11.8(h)
n.9(z(u)-)-23 cis a(s)c-3to urcoe8111ce-4123n8111c-0.3(s-313123n)3() (81113131. R3 c)-1138(o)48 sae1.4(n)-13to u81.2(pt)-34e43(m)08me43(m)08 o28.p(

PDF Overview

History of PDF

In 1985, Adobe helped create what was then called "the desktop publishing revolution" by introducing the Adobe PostScript® page description language. This allowed desktop printers to render complex text and graphics images. For the first time, any individual with a computer could accomplish high-end document publishing; no longer was it the exclusive realm of specially trained tradespeople. This was one of the killer applications that drove individuals and businesses to make the change from typewriters to personal computers.

In 1992, John Warnock, co-founder of Adobe Systems Incorporated, speaking about the goals of a development project known as Camelot, said, "There is no universal way to communicate and view this printed information electronically… What industries badly need is a universal way to communicate documents across a wide variety of machine configurations, operating systems, and communication networks."

The only attribute missing from his description was "over time." The Camelot project developed the technology known as PDF. PDF leveraged the ability of the PostScript language to render complex text and graphics and brought this feature to the screen as well as the printer.

PDF Basics

PDF is a publicly available specification, regardless of the fact that Adobe created it and advances the specification through subsequent releases. Many people confuse PDF, the data format, with Adobe Acrobat, the software suite that Adobe sells to create, view, and enhance PDF documents. In 1993, the first PDF specification was published at the same time the first Adobe Acrobat products were introduced. Since then, updated versions of the PDF specification continue to be available from Adobe via the Web. The current version of PDF specification at the date of this publication is version 1.4 and is available at *http://partners.adobe.com/asn/developer/acrosdk/docs.html*. All of the revisions for which specifications have been published are backward compatible, that is, if your computer can read version 1.4, it can also read version 1.3 and so on. Since Adobe chose to publish the PDF specification, there is an ever-growing list of creation, viewing, and manipulation tools available from other vendors.

FOR MORE INFORMATION
Two excellent sources for information regarding Adobe Acrobat and third-

The term *Portable Document Format*, or *PDF*, was coined to illustrate that a file conforming to this specification can be viewed and printed on any platform—UNIX®, Mac OS, Microsoft® Windows®, and several mobile devices as well—with the same fidelity. A PDF document is the same for any of these platforms. It consists of a sequence of pages, with each page including the text, font specifications, margins, layout, graphical elements, and background and text colors. With all of this information present, the PDF file can be imaged accurately for the screen and the printing device. It can also include other items such as metadata, hyperlinks, and form fields.

In order to ensure the specification can be used by third-party developers, Adobe has provided both an SDK and the Adobe PDF Library. Entire solutions can be developed outside of the Acrobat product family, or the Acrobat products can be modified with the development of internal plug-ins. Developers have even used just the PDF specification to create their own PDF viewers or creators. Every aspect of the file format and the manner in which it can be created, read, and manipulated is detailed in these documents. By providing this level of support, Adobe has encouraged support and use of PDF from a variety of sources.

PDF as an Archiving Format

There are many electronic formats and technologies to choose from for archiving. These include ASCII (for text), TIFF, PDF, and XML—not to mention word processing, spreadsheets, and other formats. The proprietary nature of some of these formats leads to the criticism that they cannot be guaranteed to continue for the long term. Only one of these formats is uniquely suited to ensuring display preservation over a long period of time. PDF represents not only the data contained in the document but also the exact form the document took. The file can be viewed without the originating application. In fact, ten years from now, and into the future, users will still be able to view the file exactly as it was created. With the addition of XML metadata to the PDF file, we can have both fidelity and accessibility. Because PDF is a publicly available specification, the information about the file format will always be in the public domain, making it a very attractive format to select for electronic archives. People with disabilities can also access the information using assistive technology. For instance, a visually impaired person might use a screen reader, available from vendors such as Freedom Scientific, Dolphin Oceanic, and GW Micro, to verbalize the text. This is done through embedded tags in the PDF file structure. These tags can be created automatically from the originating application or entered as part of an enhancement process.

Many organizations that are using electronic archives are implementing procedures that limit the formats of records they will receive and store. This reduces the number of file format investigations and support mechanisms that are required. The Dutch National Archives is currently supporting electronic document archive formats like PDF and XML. The Australian Victorian Electronic Record Strategy (VERS) uses XML to encapsulate PDF records along with standardized metadata. The U.K. Public Record Office limits its formats for transfer into the archives to PostScript, TIFF, SGML, and PDF.

When specifying any technology or format for employment by a broad range of users, organizations will want to ensure that they receive files in a standard manner. To ensure compliance and to prepare their user community, organizations must determine the characteristics of a well-formed PDF file that will meet an organization's archival needs. This may include limiting the use of add-ons, such as embedding multimedia or JavaScript, and insisting that the fonts used in the document be embedded or that the full scope of metadata be entered by the creator or recipient of the record (see sidebar).

There is currently an effort being spearheaded by government agencies, industry representatives, AIIM International, and NPES/CGATS to create an ISO standard around PDF. This standard would be for use specifically within the archiving community. This new project, called PDF/A, will have a broad-reaching effect on record keeping around the world. For more information on PDF/A, go to *www.aiim.org*.

Items to Watch for in PDF Archives

1. Ask that the fonts be embedded in the document.
2. Standardize on what metadata is expected in a file, and make sure users fill it in.
3. Make sure no file is submitted with passwords or encryption.
4. Discourage the use of embedded executable code.
5. Standardize the method of linking between files (e.g., use relative links if files are submitted together).

File Format and Metadata Standards

Standards: De Jure, De Facto, and Mandated

Standards that are endorsed by a standards body such as the International Organization for Standardization (ISO) or the American National Standards Institute (ANSI) are referred to as de jure standards. Data format standards that become standards by sheer volume of usage and acceptance by users are called de facto standards.

De jure standards take a long time to develop and must be approved by every organization that is a member of the standards organization with interests in the area covered. These standards bodies generally include industry members, technology developers, engineers, and specifications experts.

Standards must be clear and concise, not left to interpretation. Vague standards can cause new interoperability problems or continue the same problems that were in existence before the standards were made. Examples of de jure standards include:

- Z39.5 *Z39.50* refers to the International Standard, ISO 23950: "Information Retrieval (Z39.50): Application Serve Defi nition and Protocol Specification," and to ANSI/NISO Z39.5.
- MARC 21: MARC is the acronym for Machine-Readable Cataloging. It defines a data format that emerged from a U.S. Library of Congress-led initiative that was begun 30 years ago. MARC became USMARC in the 1980's and MARC 21 in the late 1990's. It provides the mechanism by whh computers exchange, use, and interpret bibliographic information, and its data eements make up the foundation of most library catalogs used today.
- JPEG: JPEG is a standardized image compression mechanism. JPEG stands for Joint Photographic Experts Group, the origina name of the commtee that wrote the standard.

De facto standards spring up in response to an immediate industry need. They gain in use and poularity through market dictates. They are usualy maintained by the group or business that originated them, and they have no communy review. These standards tend to be narrower in scope and designed for one specific purpose. They penetrate the market and become a standard by virtue of the fact that they solve key industry problems. PostScript and PDF are both examples of de facto standards.

Mandated standards include those that are either regulated or "suggested for compatibility." Examples of organizations that mandate standards are the U.S. Food and Drug Administration (FDA) regarding New Drug Applications (NDAs). These can be submitted electronicaly, but the FDA wants the document portion as a PDF file with very specific attributes, such as bookmarks and hyperlinks. In fact, the FDA publishes guidance documents on the subject of acceptable electronic formats at *www.fda.gov/cder/guidance*. Austraia's Victorian Electronic Records Strategy (VERS) project mandates PDF for records that are subsequently wrapped with XML records metadata. By establishing appropriate policies and procedures, individual organizations are effectively mandating standards for their own internal use.

Another example of a mandated standard is the use of PDF/X. PDF/X is an ISO standard that was initially created in the press and advertising communities. In those industries, whether an organization creates or receives PDF documents, they must be ready for output right away. This is a PDF for graphic art professionals and has a specific subset of standards that comply with the ISO standard 15930-1: 2001. There is more information about PDF/X at *www.pdfx.info*.

Another example of the implementation of a mandated standard is the use by the Administrative Office of the U.S. Courts and the U.S. Bankruptcy Court for the Southern District of New York of a Web-based application that enables lawyers to file bankruptcy documents electronically as PDF files. The public can then view these documents at *www.nysb.uscourts.gov.* Many high-profile bankruptcy cases have been filed using PDF, saving countless hours of labor and reams of paper. These filings are the actual records. There are no follow-up documents on paper. According to Cecelia Morris, a former clerk of the U.S. Bankruptcy Court for the Southern District of New York, the chief benefit of filing legal documents as PDF files is improved service to the court's constituents. In the past, accessing the files, which are a matter of public record, was difficult and time-consuming. An interested party would have to come to the court in person or wait for documents to be mailed. Now when somebody calls the court, he or she is simply referred to the Web site. The cross-platform, cross-application nature of a PDF file means that the file appears exactly like the original, no matter which platform the requester is using. "Page fidelity is a critical advantage of PDF for legal documents," says Morris. "This is the official document, and PDF ensures that every word is identical to the original."

The Role of Metadata in Electronic Archives

In order to create electronic archives, emphasis must be given to the creation of metadata. The term *metadata* originally emerged in the IT community. This concept of "data about data" has been used by information professionals to describe information about an object. Traditional records management tools such as file registers, file covers, movement cards, thesauri, and indexes all provide metadata about records. Such tools help records managers control and manage records. Additionally, records management tools provide important contextual information about who used the records, how they were used, and when they were used. In the past, archivists provided additional metadata by creating indexes, file lists, and other search aids that helped researchers locate and understand records once the records were transferred to archival custody. Today's metadata is a more proactive event with system designers trying to capture data about the data from the source, either an automated system or the actual author of the information.

Dublin Core Metadata Initiative uses Resource Description Framework (RDF) because RDF allows metadata schemes to be read by humans as well as parsed by machines and allows multiple objects to be described without specifying additional detail. The underlying glue, XML, simply requires that all namespaces be defined. Once they are defined, they can then be used to the extent needed by the provider of the metadata. Dublin Core metadata elements can be contained within PDF files. The following example of metadata extracted from a PDF file identifies the Dublin Core namespace (xmlns='http:.5708 TDt 59 -0.(o)9.2(00 (ad i)-234.9(

Organizations must establish an organizational metadata standard that will specify the type of information that will describe the identity, authenticity, content, structure, context, and essential management requirements of records. This standard, descriptive information will enable reliable, meaningful, and accessible records to be carried forward through time to satisfy business needs and evidential requirements.

There are a variety of international efforts to establish metadata standards. These efforts can provide good beginning points for an organization to consider standards in metadata practices.

- Dublin Core Metadata Initiative, *http://dublincore.org/index.shtml*, August 2002
- Victorian Electronic Records Strategy Project, "VERS Metadata Scheme: Public Record Office Standard, PROS 99/007, Specification 2," *www.prov.vic.gov.au/vers*, August 2002
- U.S. Department of Defense, "Department of Defense Directive 5015.2 Standard," *www.dtic.mil/whs/directives/corres/html/50152.htm*, August 2002
- MARC 21, *www.loc.gov/marc*, August 2002

The Archiving Process

The Workflow from Creation to Archive

The path to archiving in an organization that employs paper-based records is simple—print the document and file. Of course, this is assuming that there is a central filing room still available, with an understandable filing plan. If this is not the case, the paper archive is only accessible to the person who filed it. To move from this to a full electronic archive is not without challenges. First, a comprehensive set of policies needs to be put in place. An organization needs to have a clear understanding of what a record is and of how long it will need to be kept. Good records management practices dictate destruction when appropriate.

A good records management program must have the organizational policies and procedures in place, the individuals trained in the application of those polices, and an auditable effort in place that follows those policies and procedures.

Most existing records management policies and procedures were designed for paper records. If an organization intends to modernize and begin archiving important documents and records in PDF, it needs to update those policies and procedures to reflect the expanded use of electronic records and to ensure that an adequate legal and historical record of its decision making continues to be maintained with regard to electronic records.

To address an electronic archive, an organization needs a central record management system. This system needs to address records that are "born" electronic, those that will be converted from paper to electronic (scanned), and those that will never be electronic (due to perceived value or costs). For records that are born electronic, desktop conversion, automated processes or server-based processes can accomplish the conversion to PDF documents. Examples of software that perform these functions are Adobe Acrobat software and Adobe Acrobat Distiller® Server. To convert paper to electronic files, Adobe Acrobat Capture can be used to scan and convert paper documents to PDF documents. Using Adobe Acrobat Capture is a quick way of digitally enabling an organization's paper archives.

PDF files are very useful as an archive format because the text in a PDF file is accessible to the full-text search indexing engine available in most records management systems. Thus, the archive can be searched across its metadata and its full text. If it was necessary to find memos created between certain dates with the words *decision* and *bankruptcy* in them, it can be done. Even the paper documents that are scanned and converted to PDF documents can be made searchable by use of an optical character recognition (OCR) engine. This technology identifies the appearance of text on a page and can convert the scanned image to recognizable text.

Migrating Archives to Ensure Preservation

Certain preservation techniques, such as integrity checks and backups, are always going to be necessary to preserve any type of digital information. For electronic records, additional preservation methods

must be implemented to ensure usability across time. Two approaches to digital preservation are emulation and migration.

Emulation is the re-creation of the technical environment required to use older digital objects, for example, running a DOS program in a Microsoft Windows operating system. Migration is routinely moving the data to new hardware and software configurations. Each move must be documented and checked for completeness. Migration is all the more reliable if all electronic records conform to a limited set of standardized formats.

One of the most significant costs associated with the life cycle maintenance of an electronic archive can be the migration cost of moving a document from one version of software to another. The effort associated with this migration of the file format can be as simple as opening the document and saving it as the new format. However, experience has shown that migration is usually not this simple. Opening documents that were created in earlier software versions can create problems with the page layout, heading numbering, graphics, and so forth. Sometimes these problems are due to the software, but they can also be caused by the manner in which the user employed, or attempted to employ, a software feature. In these situations, a user may have to spend time reformatting the document to make sure that it looks identical to the original document. If there is not a hard copy of the original document available and the user does not have a working copy of the previous version of the software, then it may be impossible to reformat the document to look exactly like the original. The hidden cost in these migration efforts is the manpower needed to ensure that the record still maintains its integrity. That is why PDF is being used by many organizations as the format to store electronic records. A PDF file represents the printed page and does not change when opened, unlike a document saved in a word processing format, which can change.

Because the PDF specifi

The Future of Digital Archives

A Worldwide Initiative

Organizations around the world are working hard to develop digital archiving standards that will endure well into the fut . ive gv fi chs prominently in thechse efforts. The iv - e F - / AIS - Os - t - a - n d - ferrechtandard for electronic archiving. Adobe is committed to supporting ive F a cations is used to ib - ic fullest potential.

Because the iv - e Fs - p - e - c - i - fi T - . Tc . Tw . Td - c - a programs to access electronic archives. In fact, there ichs al . chady a robust community of dvelopers in existence today who create tools for ive F . T - h - i - s c - o m - m u - n - it -

The c - o mb i - n - at - io n of worldw - ide s - t - a - n d - a - r d - s

ive Fw - i - l - la - s - s - u - re s - t - chamlined and continuous access to archives

Resources

"The L - o ng - T e - r - m ivh - reservation of Authentic Electronic Records: Findings of *interpares* ww *book/index cfm*

Dvhgital Audio-Visual Preservation Prototyping Project
www.loc.gov/rr/mopic/avprot/avprhome.html [TJ/T1_31Tf0.0005Tc0.0005Tw02.3158TD[(U)30(K)20(Re)12(c)1

Adobe Systems Incorporated • 345 Park Avenue, San Jose, CA 95110-2704 USA • *www.adobe.com*

Adobe
Tools for the New Work™