

PDF/A – worldwide collaboration to preserve electronic documents

by Susan Sullivan,
US technical advisory group for
the PDF/A ISO standard

Just about everyone who has used the Internet has come in contact with the Portable Document Format (PDF). Adobe Systems Incorporated openly publishes the PDF specification and encourages vendors to use it to develop software that creates and processes PDF files.

PDF is a digital format for representing documents. PDF files may be created natively in PDF form, converted from other electronic formats or digitized from paper, microform, or other hard copy format. Businesses, governments, libraries, archives and other institutions and individuals around the world use PDF to represent considerable bodies of important information. Much of this information must be kept for substantial lengths of time; some must be kept permanently. These PDF files must remain useable and accessible across multiple generations of technology. The future use of, and access to, these objects depends upon maintaining their visual appearance as well as their higher-order properties (such as the logical organization of pages, sections and paragraphs) machine recoverable text stream in natural reading order, and a variety of administrative, preservation and descriptive metadata.

Ensuring accessibility over time

The problem is that the feature-rich nature of PDF can create difficulties in preserving information over the long term. For example, PDF documents are not necessarily self-contained; some files depend on system fonts and other content drawn from outside the file. As technology changes, these external dependencies can cause information to be lost. Additionally, because there are many PDF development tools on the market, there is inconsistency in the file format. This means that future migration of PDF files could be difficult because archivists won't necessarily know "what's under the hood".

With so much important information all over the world being maintained as PDF, we needed a long-term solution to ensure that digital PDF documents remain accessible for long periods of time.

Addressing different market/application needs

ISO 19005-1, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)* is the first in a new family of ISO standards to address the growing need to maintain information in electronic documents over archival time spans.

This is where the "A" in PDF/A comes in. Although never formally defined, the "A" to most of the people involved in the standards activity represents Archive. This is accomplished by identifying a limited set of PDF objects that may be used in PDF/A, and adding restrictions to the use, or form of use, of those objects and/or keys within those objects.

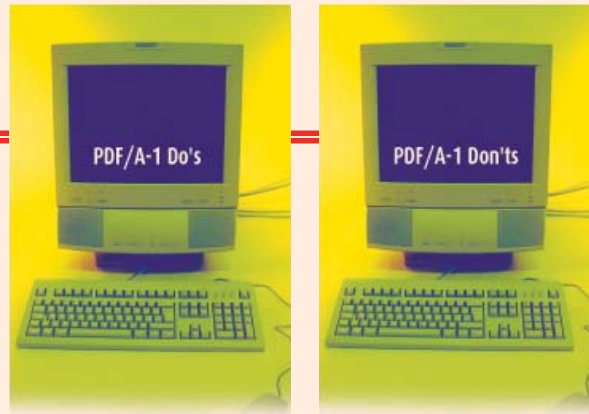
The PDF/A activity was initiated in the USA through the joint sponsorship of the Association for Information and Image Management (AIIM) and the Association for Suppliers of Printing,

Publishing, and Converting Technologies (NPES). Under the auspices of ISO technical committee ISO/TC 171, *Document management applications*, subcommittee SC 2, *Application issues*, a Joint working group (JWG 5) was formed with representatives from ISO/TC 171, ISO/TC 42, *Photography*, ISO/TC 46, *Information and documentation*, and ISO/TC 130, *Graphic technology*. A diverse group of librarians, archivists, PDF software developers, government agencies, imaging experts, graphics experts and others collaborated to develop PDF/A-1. Initial meetings were held in mid-2002 and the standard was approved in June 2005. Technical experts from 15 national standards bodies provided input throughout the development process.

“Just about everyone who has used the Internet has come in contact with the Portable Document Format.”

The PDF/A-1 (ISO 19005-1:2005) standard is based on Adobe's PDF Reference 1.4, and specifies how to use a subset of PDF components to develop software that creates, renders and otherwise processes a flavour of PDF that is more suitable for archival preservation than traditional PDF. It aims to preserve the static visual appearance of electronic documents over time and to support future access and future migration needs by providing frameworks for: (1) embedding metadata about electronic documents, and (2) defining the logical structure and semantic properties of electronic documents.

Because this initial version of PDF/A-1 is based on PDF 1.4, the standard is being published in parts so that new parts can be added without obsolescing previous parts. For example, PDF/A-1 refers to the format defined by part 1 (ISO 19005-1) of the standard (or PDF 1.4) while part 2 (ISO 19005-2) and later parts may be based on a later version of PDF and/or may define



- Embed fonts ;
- Device-independent colour ;
- XMP metadata ;
- Tagging.
- Encryption ;
- LZW Compression ;
- Embedded files ;
- External content references ;
- Transparency ;
- Multimedia ;
- JavaScript.

archiving requirements for more complex content types.

PDF/A is one of many efforts underway to build standards which use PDF as the underlying document format, but which address different market/application needs. Current standards based on PDF, either in work or already published include:

- PDF/X for pre-press data exchange ;
- PDF/E for engineering, architectural and GIS documents ;
- PDF/UA for handicapped accessibility.

An important goal is for PDF/X, E, and UA to be conformant with PDF/A and therefore capable of being preserved over the long term. Currently, it is possible for a file to be compliant with both PDF/A-1:2005 and the currently published versions of PDF/X (ISO 15930 parts 1, 3, 4, 5 and 6).

Designed for flexible implementation

For an archival standard to be viable, it must allow for flexibility of implementation. For example, organizations will want to implement the PDF/A file format at various stages of the document lifecycle, possibly even upon document creation.

The committee developing PDF/A-1 wanted to be sure that PDF/A could be used by everyone, and not just by libraries, records managers and archival institutions. We recognized that organizations would use PDF/A applications to create and process PDF/A conformant files as part of their regular business processes and, in doing so, would need to adhere to different rules and requirements.

In defining PDF/A-1 as a file format standard, the committee limited its scope to define an archival version of the PDF 1.4 file format. This would allow the flexibility needed for wide implementation, and leave to its implementers : processes for generating PDF/A-1 files, specific implementa-

tion details of rendering such files, file storing methods and hardware/software dependencies.

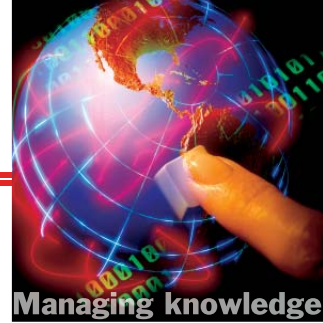
The introduction clearly explains that, as a file format standard, PDF/A-1 is just one component of an organization's comprehensive preservation strategy and does not stand alone. By itself, PDF/A-1 does not guarantee preservation or exact replication of source material. Implementers must use it in conjunction with additional controls such as : records man-

About the author



Susan Sullivan is a Certified Records Manager and Information Technology Specialist for the US National Archives and Records Administration

(NARA). She represents NARA on the US technical advisory group for the PDF/A ISO standard. For 20 years, Ms. Sullivan was a records management consultant to the nuclear power industry and then to the Federal government. In 2002, she joined NARA as a member of the ERM E-Gov Team and recently participated in the electronic records policy working group. Currently, she leads the NARA "Toolkit for Managing Electronic Records" project.



agement policies and procedures, quality assurance processes, requirements and conditions necessary to ensure persistence of electronic documents over time. For example, PDF/A-1 information annex B recommends conversion processes to ensure that PDF/A-1 files retain their quality and integrity as records.

PDF/A-1 supports two conformance levels to promote the creation of PDF/A-1 files with rich semantic and structural information, and to allow less complex files such as scanned images. There are two levels of conformance. Level A uses Tagged PDF and Unicode character maps to preserve the document's logical structure and content text stream in natural reading order, and level B includes all requirements of ISO 19005-1 minimally necessary to preserve the visual appearance. While level A should support a higher level of document preservation service and confidence over time, level B allows PDF/A-1 conformance without requiring users to define structure or other descriptive information.

Meeting long-term preservation needs

The PDF/A Joint working group identified desirable properties for a long-term preservation format. We adopted them as our objectives in developing the PDF/A-1 standard to ensure that it would meet long-term preservation needs.

Our intent was not to claim that PDF-based solutions are the best way to preserve electronic documents. We simply defined PDF/A-1 as an archival profile of PDF that is more amenable to long-term preservation than traditional PDF.

An archival profile of PDF

The PDF/A-1 standard is organized to mirror the PDF 1.4 Reference. In developing the standard, we assigned each section to a subcommittee, composed of PDF developers, archivists, librarians, and content managers. These experts collaboratively evaluated each section of the PDF Reference against desired properties of a long-term format (i.e. our objectives). Based on this evaluation we identified PDF com-

ponents that could complicate archival preservation and prohibited or restricted them. For each section, we defined the set of PDF objects that may be used in PDF/A-1 and in many cases added restrictions to their use. We also specified PDF objects that may not be used in PDF/A-1 files. In some cases, we defined how a PDF/A-1 conformant reader must handle these objects.

Let's look at how PDF/A-1 addresses long-term preservation needs:

- **Device independence** – PDF/A-1 requires device independent components so that the static visual appearance can be reliably and consistently rendered and printed without regard for the hardware or software platform used. The graphics clause, for example, incorporates requirements from PDF/X to ensure predictable colour rendering. PDF/A-1 also prohibits the use of components not defined in PDF Reference 1.4.
- **Self-containment** – everything that is necessary to render or print a PDF/A-1 file must be contained within the file. The fonts clause requires that all fonts used are embedded in the file. A PDF/A-1 conforming writer must always embed fonts, meaning that the file will be rendered using the fonts intended and not those residing on the local workstation. A conforming reader must always use the embedded fonts. The standard warns creators that not all fonts can be legally embedded and that legal restrictions on embedding should be determined.

The annotations clause prohibits embedded files because such files rely on external software for rendering. In the future, some software programmes could be unavailable, and the information within embedded files could be lost.

- **Self-describing files** – PDF/A-1 requires the Adobe Extensible Metadata Platform (XMP) be used for embedding metadata in PDF files. Again, to allow flexibility of implementation, PDF/A-1 provides recommendations for documenting file attributes (such as file identifier, file provenance, font meta-

data), and allows non-XMP schemas to be included, as long as they are embedded. Implementers can use XMP in a variety of ways to include information about electronic records within the file itself. Having metadata embedded in the file can increase the informational value of electronic documents and enhance the future researcher's understanding of the document.

“For an archival standard to be viable, it must allow for flexibility of implementation.”

- **Transparency** – level A conforming PDF/A-1 files provide text “in natural reading order” so that the file can be read with basic text editing tools, such as MS Notepad. This supports access to the informational content of PDF/A-1 files, even without the benefit of a PDF/A-1 reader.
- **Accessibility** – PDF/A-1 prohibits encryption in the file trailer. This prohibition means that user IDs and/or passwords are not needed to do anything with a PDF/A-1 file. PDF/A-1 files are open and available to anyone or any software that processes the file. Implementers that require access controls can provide them outside of the file format.
- **Disclosure** – PDF/A-1 is based on an authoritative specification that is publicly available. Anyone can use the PDF reference and XMP specification in conjunction with PDF/A-1 to create applications that read, write, or process PDF/A-1 files. Adobe has granted a general royalty free license to use certain of its patents to create applications that process PDF/A files. Additionally, it has granted AIIM and NPES the rights to publish these specifications on their respective Internet sites for the foreseeable future.
- **Adoption** – PDF/A-1 was designed for flexibility of implementation to

promote its wide adoption. If widely adopted, PDF/A-1 software tools will proliferate and the market will support the file format, to help ensure the viability of PDF/A-1 and extend the length of time that PDF documents can be maintained as PDF/A-1, as long as the demand exists.

Reliability and predictability

The requirements of PDF/A-1 emphasize reliable and predictable rendering of static visual appearance. Users should use a PDF/A-1 conformant viewer to view or print PDF/A-1 documents. Conformant viewers should ensure reliable visual representation of the document. Also, PDF/A-1 permits the inclusion of interactive elements (e.g. annotations and hyperlinks) but suggests that a conformant viewer treat them as inactive. Implementers should take these needs into account when choosing a conformant viewing tool.

PDF/A may not be the last preservation format you will need, but proper application of it should result in reliable, predictable and unambiguous access to the full information content of electronic documents.

Work has already begun on PDF/A part 2 (PDF/A-2) and part 3 (PDF/A-3) which will be based on PDF 1.6 (which subsumes PDF 1.5). Implementers have requested that the following features be considered for inclusion in future parts of PDF/A:

- JPEG 2000 image compression;
- More sophisticated digital signature support;
- Open type fonts;
- 3D graphics;
- Audio/video content;
- Consistency with PDF/X, PDF/E and PDF/UA.

The PDF/A Joint working group is creating both a set of application notes and a list of frequently asked questions which will be made publicly available to assist developers of PDF/A applications to better understand the requirements of the file format and provide implementation guidance. ■



The learning curve – how IT shapes learning environments

by Bruce Peoples, Chair of ISO/IEC JTC 1/SC 36, Information technology for learning, education and training

In the domain of learning, education, and training, knowledge management within information technology-based learning environments requires specific standards to supplement existing and emerging knowledge management standards.

With a membership of 28 national standards bodies and 20 liaison organizations, ISO/IEC Joint technical committee JTC 1, *Information technology*, subcommittee SC 36, *Information technology for learning, education and training*, serves as the pre-eminent international forum for standards develop-

ment in IT for learning, education, and training. In fulfilling a leadership role in the standards domain, the subcommittee has an emerging reputation for quality and innovation – by utilizing proven, value-adding processes and procedures in producing International Standards and technical reports. Only by ensuring quality and innovation for the implementers and users of SC 36 documents, will the global use of IT in the creation of learning environments achieve its potential.

Learning environments are unique in that standardization and interoperability are required across many levels of implementation, including the low-end (e.g. a standalone PC workstation, or Personal Digital Assistant), the middle (e.g. workstations with high-speed internet access) and the high-end (e.g. high-fidelity simulators and trainers).

Learning environments consisting of e-enabled learning services across the globe are also unique in the realm of IT and standardization because of their embedded cultural attributes derived from the unique educational and social heritage of each region, nation and language.