Standards Update

David Q. McDowell, Editor

On October 1, 2005, ISO published a standard called ISO 19005-1, *Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1)*. This event was far more momentous than it appears at first blush.

Part of the significance of this event is the cooperative way in which this document was developed, much of the rest lies in the interest, and support, of the US and the international government and archival community.

Lets first review how this document was brought into existence, then we will look at the requirements of ISO 19005-1, and then the plans for the follow-on work.

How it came to be

In the Spring of 2002 several different groups (The Administrative Office of U.S. Courts (AOUSC), Library of Congress, IRS, etc.) realized that PDF was becoming significant as a dominant form for electronic document storage and exchange.

A group was formed under the sponsorship of AIIM (Association for Information and Image Management) and NPES (The Association for Suppliers of Printing, Publishing, and Converting Technologies) to propose ways that could ensure preservation of PDF documents over extended periods of time, and to further ensure that PDF documents would be rendered with consistent and predictable results in the future.

AIIM brought to the table a long history of document storage and retrieval. NPES, as secretariat of ANSI/CGATS and TC130/WG2, had already been the leader in the development of PDF/X—the application standard that defines the use of PDF for graphic arts data exchange. The government agencies and archival community brought an understanding of both the user needs and the archival requirements.

This new group became known as PDF/A where the A became commonly accepted as "for Archiving".

Problem Description

The problem description created out of this group had as its opening statement the following:

"Customers who intend to archive PDF documents have unique requirements to ensure that the intended document is readable for an indefinite amount of time.

In order to allow PDF to meet this requirement they believe that a subset of PDF should be published and controlled by an external group who would be capable of overseeing it's evolution to ensure its longevity.

Additionally, customers need a means to defined business rules for archives and allow for simplified creation, validation and processing of the archived collection."

Getting ISO Involvement

Based on the initial work of the PDF/A group, AIIM initiated a New Work Item (NWI) proposal in ISO TC 171/SC2 which seemed to be the logical home for such an activity. However, it was obvious from the start that other groups had a stake in the outcome of such a standard.

The NWI was approved and TC171/SC2 established a Joint Working Group (JWG) and invited all interested ISO Technical Committees to participate.

Currently the following are members of this JWG: TC171/SC2 (Document management applications-Application issues-PDF/A); ISO/TC 42, Photography; ISO/TC130, Graphic technology; and ISO/TC 46/SC 11, Information and documentation—Archives/records management.

The first document produced by this JWG is ISO 19005-1, *Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1).*

Why PDF and what are we trying to accomplish?

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.

Adobe Systems Incorporated makes the PDF specification publicly available. However, the inclusive, feature-rich nature of the format requires that additional constraints be placed on its use to make it suitable for the long-term preservation of electronic documents.

The primary purpose of ISO 19005 is to define a file format based on PDF, known as PDF/A, which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files.

A secondary purpose of ISO 19005 is to provide a framework for recording the context and history of electronic documents in metadata within conforming files.

A third purpose is to define a framework for representing the logical structure and other semantic information of electronic documents within conforming files.

These goals are accomplished by identifying the set of PDF components that may be used, and restrictions on the form of their use, within conforming PDF/A files.

Some characteristics of PDF/A

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 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 that in doing so, they would need to adhere to different rules and requirements.

In defining PDF/A-1 as a file format standard, we limited PDF/A-1's scope to define an archival version of the PDF 1.4

file format. This would allow the flexibility needed for wide implementation.

PDF/A-1 supports two conformance levels to promote the creation of PDF/A-1 files with rich semantic and structural information, but also to allow less complex files such as scanned images. The two levels of conformance are referred to as Level A and Level B. Level A provides a higher level of structural and descriptive information, and Level B includes all requirements of 19005-1 minimally necessary to preserve the visual appearance. Level B was necessary to allow PDF/A conformance without requiring users to define structure or other descriptive information.

PDF/A-1 leaves to its implementers: processes for generating PDF/A-1 files, specific implementation details of rendering PDF/A-1 files, methods for storing PDF/A-1 files, and hardware/software dependencies.

Implementers will need to use additional controls to ensure quality, authenticity and integrity of PDF/A-1 documents including quality assurance processes to validate replication of source material.

The Scope and Introduction point out that, as a file format standard, PDF/A is one part of an organization's long term archival environment and does not stand alone.

What else is needed

By itself, PDF/A-1 does not necessarily ensure that the visual appearance of the content accurately reflects any original source material used to create the conforming file; e.g. the process used to create a conforming file might substitute fonts, reflow text, downsample images or use lossy compression.

Organizations that need to ensure that a conforming file is an accurate representation of original source material may need to impose additional requirements on the processes that generate the conforming file beyond those imposed by ISO 19005-1.

In addition, it is important for those organizations to implement policies and practices regarding the inspection of conforming files for correct visual appearance.

ISO 19005-1 should be used as one component of an organization's electronic archival environment for long-term retention of documents.

Successful implementation of t ISO 19005-1 for archival purposes depends upon:

- the retention requirements of an organization's archival environment, records management policies and procedures as specified in ISO 15489-1:2001, Information and documentation — Records management — Part 1: General;
- any additional requirements and conditions necessary to ensure the persistence of electronic documents and their characteristics over time, including, but not limited to, those defined by: ISO 14721, *Space data and information transfer systems—Open archival information system—Reference model*; ISO/TR 15801, *Electronic imaging — Information stored electronically—Recommendations for trustworthiness and reliability*; ISO/TR 18492, *Long-term preservation of electronic document-based information*; ISO 18509-1, *Electronic archival storage—Specifications relative to the design and operation of information processing systems in view of ensuring the storage and integrity on recordings stored in these systems—Part 1: Long term access strategy*; ISO 18509-2, *Electronic archival storage—Specifications relative to the design and operation of information processing systems in view of ensuring the storage and integrity on recordings stored in these systems—Part 2: Technical specifications quality assurance processes necessary to verify conformance with applicable requirements and conditions*; e.g. an inspection regime to verify the quality and integrity of converted source data.

Application development

ISO 19005-1 will lead to the development of various applications that read, render, write and validate conforming files.

Different applications will incorporate various capabilities to prepare, interpret and process conforming files based on needs as perceived by the suppliers of those applications. However, it is important to note that a conforming application must be able to read and process appropriately all files complying with a specified conformance level.

Future development

This first PDF/A standard has been created as Part 1 of ISO 19005 to allow the creation of future parts These parts can provide compatibility with future versions of the underlying PDF specification without rendering this document or applications based on PDF Version 1.4 obsolete and/or can provide more complex or richer electronic document storage capabilities as new technology becomes available.

Work has already begun on PDF/A Part 2 which will probably be based on PDF 1.6 (This subsumes PDF 1.5).

Implementers and users have requested that the following features be considered for inclusion in Part 2.

- JPEG 2000 image compression
- More sophisticated digital signature support
- Open Type fonts
- 3D graphics
- PDF Transparency
- Layers (optional content)
- Audio/video content
- Consistency with PDF/X, PDF/E, PDF/UA

In addition the PDF/A JWG is creating both a set of Application Notes and a listing of Frequently Asked Questions (FAQs) 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.

Copies of ISO 19005-1:2005, PDF/A-1, may be purchased from ISO (www.iso.org), or in the US from NPES (http://www.npes.org/standards/orderform.html), or any National Standards office.

Much of the material for this article came directly from the standard itself or from various documents and reports prepared by members of the NPES/AIIM PDF/A Committee.

———————————————————

For suggestions for (or input to) future updates, or standards questions in general, please contact the author at mcdowell@npes.org or mcdowell@kodak.com