
# FORMATS FOR DIGITAL PRESERVATION: A REVIEW OF ALTERNATIVES AND ISSUES

*Submitted By*

**CENDI Digital Preservation Task Group**

***Revised***

**March 1, 2007**

***CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Agriculture, Commerce, Energy, Education, Defense, the Environmental Protection Agency, Health and Human Services, Interior, the National Aeronautics and Space Administration, the Government Printing Office, the National Archives and Records Administration, the National Science Foundation, and the Library of Congress.***

***CENDI's mission is to help improve the productivity of federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.***

# TABLE OF CONTENTS

# EXECUTIVE SUMMARY

CENDI Members recognize digital formats as acceptable means of preserving Government information (CENDI, 2007). This review of alternative formats and the issues related to them was undertaken in the interest of implementing best practices in information life-cycle management, to dispel any misunderstandings related to digital formats, and to provide agencies with enough information so they can determine what the most appropriate preservation format is for them.

## BACKGROUND

CENDI Members' interest in digital preservation formats was spurred on by the Office of Management and Budget's (OMB) call for government information management standards, the Chief Information Officers (CIO) Council's response to the E-government Act, the identification of archival formats for the digital deposit or records, and the development of agency repositories. In 2005 the CENDI Members requested an assessment of the digital formats being used for preservation and the issues surrounding them. The CENDI Digital Preservation Task Group submitted their final report on December 22, 2006.

Many digital file formats can be considered for preservation. CENDI agencies, however, are most concerned with formats that best preserve text documents such as technical reports and journal articles. For this reason the report focuses on four major formats in the context of document preservation – TIFF, PDF, PDF/A, and XML.

## FORMAT ASSESSMENT FACTORS

The appropriateness of TIFF, PDF, PDF/A, and XML formats was assessed by the Library of Congress as part of a more comprehensive evaluation using the following:

- **Technical Factors**, each format is analyzed against the following factors for sustainability:
  - Disclosure – existence of complete documentation
  - Adoption – degree to which the format is already in use
  - Transparency – degree to which the digital representation is open to direct analysis
  - Self-documentation – digital objects that contain basic descriptive, technical, and other administrative metadata
  - External Dependencies – degree to which the format is dependent upon specific hardware, operating system, or software for rendering or use and the complexity of dealing with those dependencies in future technical environments
  - Impact of Patents – degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents
  - Technical Protection Mechanisms – implementation of mechanisms that prevent the preservation of content by a trusted authority
- **Quality and Functionality**, importance of content for reuse versus layout and presentation
- **Striking a Balance**, between the technical factors and the quality and functionality factors, which may compete with one another and may change over time

## ASSESSMENT RESULTS

The results of the Library of Congress's assessment of the formats of interest in this white paper are summarized in the following tables (Arms & Fleischhauer, 2006). Table 1 discusses each format against the sustainability factors. Table 2 summarizes LC's findings for each format against the criteria related to quality and functionality.

| SUSTAINABILITY FACTORS | FILE FORMATS | | | |
|---|---|---|---|---|
| | **PDF** | **PDF/A** | **XML** | **TIFF_G4** |
| **DISCLOSURE** | Fully documented. PDF was developed by Adobe Systems Incorporated, which makes the specification available openly and at no charge. One subtype of this proprietary format has been adopted as an international standard by ISO (PDF/X). A second is in the standardization process (PDF/A). | Open standard, approved in May 2005 and published by ISO in September 2005. Developed by the working group ISO/TC 171 SC2, Document Imaging Applications, Application Issues, for which AIIM (The Association for Information and Image Management) acts as secretariat. ISO has formed a Joint Working Group, which also includes ISO/TC 46 SC11, Archives/records Management, ISO/TC 130, Graphics Technology, and ISO/TC 42, Photography. | Open standard. Developed by World Wide Web Consortium. To be useful for interoperability or long-term content preservation, an XML document must be associated with a schema specification for the elements and tags it contains. Such schema specifications must also be disclosed. | Fully documented. TIFF was developed by the Aldus and Microsoft Corporations, and the specification is owned by Aldus (now absorbed into the Adobe Corporation). The TIFF tag set is extensible through a registry maintained by Adobe; the list of registered extensions is not available from Adobe; see Tags for TIFF and Related Specifications. |
| **ADOPTION** | Extremely widely adopted as a platform-independent format for disseminating page-oriented documents. Adobe Reader software for viewing PDF files is freely distributed and bundled with most personal computers. | Tools for creating, converting, and validating have reached the market steadily since the standard was published in 2005. Acrobat Professional 7.0 allows saving files in a form compliant with the draft standard. Acrobat 8 supports the standard as published. During 2006, several commercial companies produced products supporting the creation, migration, and validation of PDF/A files. The growing requirements from the EU for use of digital formats that are formal (preferably ISO) standards has produced more market pressure than in the U.S. Version 0.93 of the widely used open source FOP (Formatting Object Processor, based on the W3C's XSL-FO standard) from Apache (released in January 2007), has support for the minimal PDF/A profile, PDF/A-1b.<br><br>The standards development process involved active participation of communities whose endorsement or adoption would create significant momentum for wider adoption of PDF/A over generic PDF for archival deposit or submission. Adobe reported migration of legacy "report silos" in November 2006. | Very widely adopted as the basis for interchange of documents and data over the Web. Many generic tools exist, including free and open source software. Major software vendors have all incorporated support for XML in some form. | TIFF_G4 is widely deployed in digital library projects as a master format and, in December 2005, the Government Printing Office (GPO) announced that TIFF_G4 had been selected as the master format for bitonal preservation images. Not supported by all browsers in native format, but, as of early 2004, new PC configurations tend to include a viewer. TIFF_G4 is acceptable for raster images in the list of FCLA recommended formats (Florida Center for Library Automation; www.fcla.edu/digitalArchive/pdfs/recFormats.pdf). |
| **TRANSPARENCY** | Depends upon compliant software tools to read. Building tools requires sophistication. | Depends upon compliant software tools to read. Building tools requires sophistication. PDF/A does not permit encryption. | Human-readable and designed for automatic parsing. A well-documented DTD, XML Schema, or other specification is needed. Human-comprehensible element tags are advantageous. | Depends upon algorithms and tools to read; requires sophistication to build tools. |
| | | | | |

| SUSTAINABILITY FACTORS | FILE FORMATS | | | |
|---|---|---|---|---|
| | **PDF** | **PDF/A** | **XML** | **TIFF_G4** |
| **SELF-DOCUMENTATION** | Later versions of PDF can include XMP metadata packages. | Support for embedding any form of metadata for a document is extremely good. Use of XMP is mandatory for basic descriptive and identifying metadata. Other XMP metadata packages can be embedded. | XML is widely used as a syntax for metadata, and metadata for all purposes can be embedded in XML documents with appropriate schema specifications. | The TIFF specification defines a framework for an Image File Header (IFH), Image File Directories (IFDs), and associated bitmaps. Each IFD and its associated bitmap are sometimes called a *TIFF subfile*. There is no limit to the number of subfiles a TIFF image file may contain. Each IFD contains one or more data structures called *tags*, each one of which is a 12-byte record that contains a specific piece of information about the bitmapped data. The TIFF specification defines a number of tags and a set of rules for extensibility; see Tags for TIFF and Related Specifications. Tags are always found in contiguous groups within each IFD. |
| **EXTERNAL DEPENDENCIES** | Faithful rendering requires that fonts be embedded. | PDF/A is constrained to avoid external dependencies. All necessary fonts must be embedded. | None | None |
| **IMPACT OF PATENTS** | Adobe has a number of patents covering technology that is disclosed in the Portable Document Format (PDF) Specification, version 1.3 and later. Adobe Reader displays additional patent numbers on launch. | Not expected to be a problem, but not investigated at this time. The standard includes ISO boilerplate text indicating "the possibility that some of the elements of this document may be the subject of patent rights." | None | None |
| **TECHNICAL PROTECTION MECHANISMS** | The PDF format offers several forms of technical protection, including encryption, that would prevent custodians of digital content ensuring accessibility in future technological environments. | PDF/A does not permit encryption. | None | None |

**Table 1: Sustainability Factors for PDF, PDF/A, XML, and TIFF Formats**

| QUALITY & FUNCTIONALITY | FILE FORMATS | | | |
|---|---|---|---|---|
| | **PDF** | **PDF/A** | **XML** | **TIFF_G4** |
| **NORMAL RENDERING** | Good support is possible, but not guaranteed. The PDF format allow creators to disallow printing and extraction of text for quotations. PDF can also be used to create documents from scanned page images; such files do not necessarily support indexing of the document text. | Good support is possible, but not guaranteed. The PDF/A format does not preclude creating documents from scanned page images; such files do not necessarily support indexing of the document text or extraction of text for quotation. | XML can represent all UNICODE characters, with UTF-8 being the default character encoding. XML tagging offers potential for explicitly representing logical structure of text, such as paragraphs and headings, and character emphasis (bold, italics, etc.). Effective support for normal rendering is dependent on an appropriate DTD or schema specification. | Good support. |
| **INTEGRITY OF STRUCTURE** | The logical structure of a document is only represented in a PDF file if the creator or process during creation takes steps to incorporate structural tagging. | The logical structure of a document is only represented in a PDF/A file if the creator or process during creation takes steps to incorporate structural tagging. The PDF/A standard recommends the representation of structural hierarchy | XML is ideal for representing document structure. | Not applicable |
| **INTEGRITY OF LAYOUT** | PDF is designed to represent the layout of page-oriented documents. | PDF is designed to represent the layout of page-oriented documents. | For textual content, best practice is to have the XML represent the logical document structure and use stylesheets to render the text in a form appropriate for the end user. | Not applicable |
| **INTEGRITY OF RENDERING OF EQUATIONS** | Can be represented by embedded graphics. | Can be represented by embedded graphics. | Requires specialized markup (e.g., MathML) and corresponding rendering engine. Scholars in many scientific disciplines are not satisfied with the performance of such rendering engines. | Not applicable |
| **BEYOND NORMAL RENDERING** | Supports embedding of media objects (in binary format) and links to external media objects, such as images, audio, or video. | Annotations may be embedded. Bookmarks may be provided. | Depends on particular DTD or schema specification. | Multi-page files supported for a sequence of images. |
| **CLARITY (SUPPORT FOR HIGH IMAGE RESOLUTION)** | Not applicable | Not applicable | Not applicable | Excellent support for images with very high spatial resolution. The standard is flexible as to color space and bit depth. In practice, 8-bit grayscale and 24-bit RGB color are common; some activities |

| QUALITY & FUNCTIONALITY | FILE FORMATS | | | |
|---|---|---|---|---|
| | PDF | PDF/A | XML | TIFF_G4 |
| | | | | create files with greater than 8 bits per channel (color or greyscale).TIFF_G4 is limited to bitonal (pure black and white) images. |
| SUPPORT FOR GRAPHIC EFFECTS AND TYPOGRAPHY | Not applicable | Not applicable | Not applicable | No support for vector graphics. |
| COLOR MAINTENANCE | Not applicable | Not applicable | Not applicable | The TIFF tag for the ICC profile (tag 34675, *InterColourProfile*) for a capture device has been added as a "private" extension in the TIFF/IT and TIFF/EP standards. Extended tags of this kind may be used in any TIFF_6 file, although they may not be recognized by all readers. ICC Profile version 4.2.0.0 (Specification ICC.1:2004-10, page 69) provides guidance for embedding ICC profiles in TIFF files: "as a single TIFF field or Image File Directory (IFD)." Meanwhile, Adobe Photoshop software appears to provide an alternate means to embed an ICC profile in a TIFF file; the compilers of this Web site seek explanatory comments from readers: how proprietary or interoperable is PhotoShop embedding of ICC profiles?<br><br>Color space is indicated in *Photometric Interpretation* (tag 262); in TIFF_6, this tag does not include sRGB as a value, although sRGB images may be delivered tagged as RGB. |

**Table 2: Quality & Functionality Factors for PDF, PDF/A, XML, and TIFF Formats**

## CONCLUSION

An agency must clearly define the purpose and the requirements for preservation and the purpose and requirements for the preservation format. Many agencies find it appropriate to store multiple digital formats for preservation. One format is used to preserve the content for reuse while another is used to preserve the original layout and presentation. A multi-format approach is more likely to support migration to more robust formats in the future.

Several factors must be weighed to determine the most appropriate digital format(s) for preserving its information. Which format is chosen depends upon the mission of the agency, the kind of information being preserved, the source and native format of the material, future uses of the digital objects, the expectations of current and future users, and how far into the future the objects are intended to remain useful. The decision regarding the most appropriate format must be made within a framework that balances the technical, quality and functionality factors as well as policy decisions, the publication process of the material to be preserved, and cost factors. The preservation format that provides this balance may change over time as new formats are adopted for creation and use.

## 1.0 Introduction

At the 2005 CENDI Planning Meeting, the CENDI Members requested an assessment of the current formats being used for preservation and the issues surrounding them. The identification of archival formats for the deposit of records, the development of agency repositories, and the call for government information management standards on the part of the Office of Management and Budget (OMB) and the CIO (Chief Information Officers) Council in response to the E-government Act spurred interest in preservation format options. Concerns were raised that PDF/A-1, in particular, might be promoted as a standard within the government, since PDF/A-1 has been discussed in many venues as the preservation format of choice. While people perceive PDF/A-1 as the panacea for electronic document preservation, federal officials should understand that there are viable options to PDF/A-1, and what to consider when selecting the best preservation format for their information and their situation. It is this concern with the misunderstanding of the place of PDF/A-1 in the scheme of preservation formats and an interest in monitoring and implementing best practices in information management that have led to this CENDI white paper.

The preservation format issue is often stated in terms of the "best" format. Based on the input from CENDI agencies, the review of the literature, and the in-depth LC-NDIIPP (National Digital Information Infrastructure and Preservation Program) assessment and framework, the question should be "What is the most appropriate format?"

## 2.0 What is a Preservation Format?

Preservation is defined as the activities required to keep materials in usable form for a long period of time. Generally, the activities discussed in the context of scientific and technical information are identified as "long-term preservation". Long-term has no specific time limit; it is long enough to be concerned about changes in technology and changes in the user community.

What is a format? "Format" is defined by the Global Registry of Digital Formats as "… a fixed, byte-serialized encoding of an information model." (Global, 2006) The LC-NDIIPP format sustainability assessment defines a format as "packages of information that can be stored as data files or sent via network as data streams (aka bitstreams, byte streams)." (Arms & Fleischhauer, 2006 – Formats, Evaluation Factors and Relationships)

Preservation formats are those file formats that provide the best chance to achieve preservation, including the ability to capture the material into the archive and render and disseminate the information now and in the future. In some cases, this may be only a few years, while in other cases it may be for the life of the republic.

Since the ability of these formats to address the needs of preservation is "in the eye of the beholder", the NDIIPP program has chosen the phrase "Sustainable Format." The Digital Preservation web site contains a list of seven factors which the NDIIPP program uses to evaluate the sustainability of any given format. (Arms & Fleischhauer, 2006)

**Disclosure.** Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can

be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation.

**Adoption.** Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems.

**Transparency.** Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor.

**Self-documentation.** Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.

**External Dependencies.** Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.

**Impact of Patents.** Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.

**Technical Protection Mechanisms**. Implementation of mechanisms such as encryption that prevent the preservation of content by a trusted repository.

## 3.0 The Major Formats

Many digital file formats can be considered for preservation as evidenced by the number of formats described in the Global Registry of Digital Formats (Global, 2006) and the number evaluated by the LC-NDIIPP assessment (Arms & Fleischhauer, 2006). However, CENDI agencies are historically concerned with a more limited number of formats, with an emphasis on the preservation of text documents, including journal articles and technical reports. (However, it should be noted that as non-text formats increase, this emphasis may change. For example, GPO noted that as digital imagery expands in quality, size and application, there is a greater need for image compression with flexibility and efficient interchange. JPEG2000 (with the file extension .JP2) delivers more efficient compression as well as features not available in previous image standards. As a preservation option for images, it has gained popularity in recent months. (Davis, 2006)

This white paper focuses on four major digital formats that have been discussed in the context of document preservation -- TIFF, PDF, PDF/A-1 and XML. This section briefly describes each of these formats. More detail, particularly an assessment of the use in scientific and technical information is provided in the following sections.

### *3.1 TIFF*

TIFF is one of the earliest formats used to preserve materials electronically. TIFF is a wrapper format capable of containing various image bitmaps, pixel-by-pixel representations of scanned pages or

pictures. One of the most common TIFF bitmaps is a bitonal (pure black and white) document image. Such images are produced by scanners and have been used to reproduce documents from at least the 1980s forward. The current specification is for TIFF version 6, though many software applications still produce TIFF version 5. Versions 5 and 6 are very compatible. The TIFF bitonal bitmaps are generally compressed with one of the algorithms developed for FAX transmission. Files formatted in this way are usually referred to by the shorthand TIFF Group 3 or TIFF Group 4. TIFF images faithfully reproduce the scanned page, but the text cannot be searched or manipulated. Adobe Systems Incorporated owns and publishes for open use the TIFF file format specification in the same manner as it owns and publishes the PDF format specification.

### *3.2 PDF (Portable Document Format)*

PDF was originally based on Postscript to make it possible to print across a variety of computers and printers. Adobe enhanced the technology so that it would provide the look and feel of a document across platforms.

### *3.3 PDF/A-1 (Portable Document Format/Archival)*

PDF/A-1 is a published International Standards Organization (ISO) standard. It is a specification or set of rules for what should NOT be included in a PDF 1.4 file in order to be able to read it later and what is allowed or required in PDF/A-1 and how to implement those objects. This specification can be implemented by Adobe and other vendors.

### *3.4 XML (Extensible Markup Language)*

XML (Extensible Mark-up Language) is an ASCII-based format that includes tags to accommodate both the mark-up of the meaning of fields and the display of the information. Using either DTDs or schema, XML requires declaration of the structure so that the information is more portable and interoperable.

## 4.0 History of the Discussion

CENDI's previous assessments of the state of the art and practice in digital preservation in 1999 and again in 2004 found the issue of preservation formats to be a major area of research and ongoing discussion.

### *4.1 Status in 1999*

The 1999 report found that those working during the early stages of archiving and preservation were faced with a large number of formats, primarily textual. (Carroll & Hodge, 1999) The number of formats had decreased primarily due to market forces that reduced the number of major players in the PC software market. For example, the Department of Energy's Office of Scientific and Technical Information limited the input formats when it first began accepting digital materials; in the environment at that time, it was difficult to gain support for the standardization of word processing packages. However, by the late 1990s, documents were being received in only a few formats, SGML (and its relatives HTML and XML), PDF (normal and image), WordPerfect and Word. Bitmapped

images, usually in bitonal form, were received wrapped in TIFF (with Group 3 or Group 4 compression) or in PDF (image).

Alternatively, some organizations accepted a variety of input formats and then transformed them for archive and preservation purposes. The American Astrophysical Society (AAS) and the American Chemical Society transform the incoming files from LaTex, Word, or WordPerfect to an SGML-tagged ASCII file. "The electronic master copy, if done well, [was] able to serve as the robust electronic archival copy. Such a well-tagged copy [could] be updated periodically, at very little cost, to take advantage of advances in both technology and standards. The content remains unchanged, but the public electronic version can be updated to remain compatible with the advances in browsers and other access technology." (Boyce, 1997)

The 1999 report also discussed the issue of retaining the look and feel of journal articles in particular. The majority of the projects reviewed used either image files - TIFF, PDF, or HTML. TIFF was the most prevalent format for those organizations involved in any way with the conversion of paper backfiles. For purely electronic documents, PDF was the most prevalent, particularly for less formal publication processes such as grey literature, theses and dissertations. At that time, the Royal Institute of Sweden Library transformed dissertations received in formats other than PDF to PDF and HTML. It was also prevalent as a distribution format among more formal publications.

Even by 1999, the early concerns about the impact of the proprietary nature of PDF on long-term preservation had begun to subside. The 1999 report states that "there appears to be little concern within the publishing community at this time. The main impetus is less likely to be its acceptability as an archival format as that it retains the look and feel of the original, can be produced and read easily by freeware products, and has a variety of tools available at modest costs that allow for full text searching. Hypertext links are also maintained, which is not true of TIFF images." However, despite the increased acceptance of PDF within the publishing community, concern remained among the national libraries and archives about its appropriateness for long-term preservation.

### *4.2 Status in 2004*

By 2004, many aspects of digital preservation had matured, including the roles and responsibilities of publishers, libraries and third-party archives, particularly for journal material. There were significantly more operational systems, including some commercially available vendor systems that could provide infrastructure. In addition, the whole area of institutional digital repositories had greatly expanded based on the work of MIT, Harvard and Cornell on infrastructures such as DSpace and Fedora.

However, despite these advances, the report found a continued concern about the appropriate preservation format(s). "The best format for long-term preservation remains elusive, perhaps because there is no single answer to the question. Instead it depends on the format type of the original object, the characteristics of the original that the preserving organization considers to be most important to preserve, and the expected use/re-use of the object in the future (e.g., distance education versus legal evidence). Most experts agree that the best format for preservation is that which is least proprietary while conveying significant aspects of the original." (Hodge & Frangakis, 2004)

In 2004, the most common formats for storing text were XML (ASCII, with or without Unicode), PDF, and TIFF. For scientific and technical text, as well as other objects, ASCII was considered the most open format, accommodating virtually all software or browsers. However, for some digital objects, ASCII was viewed as problematic when paired with the requirement to provide permanent access and to render the look and feel of the original. Therefore, PubMed Central, the DiVA Academic Archive Online Project at Uppsala University and the Royal Technology Library of Sweden, and the Humboldt University in Germany cited XML as the preferred format for preservation. This preference resulted from the fact that XML is based on ASCII, is non-proprietary and is well-adapted for re-purposing and interoperability. The PubMed Central Guidelines required separate SGML or XML files for the full text of each article. DiVA created XML for all available full text and Unicode was used to preserve the extended character sets from the original.

TIFF, an image format, was used to preserve the look and feel of original text objects. The use of TIFF in text environments began with the advent of scanning and Optical Character Recognition technologies, which used the TIFF images. TIFF can be employed at various resolutions depending on the quality and flexibility of the equipment used and the requirements for future use of the archived objects.

At that point, TIFF was increasingly giving way to PDF, as more capture systems supported the creation of PDF from the TIFF images. In addition, PDF was more readily created from existing authoring tools, was often the preferred choice for submission by authors, had viewers that were becoming more ubiquitous, and was more easily and reliably indexed for full-text searching. While some organizations surveyed and interviewed for the report cited issues with PDF's proprietary, though openly documented nature, PDF appeared to have gained acceptance in many quarters. For some organizations, this was probably a pragmatic move, since it is possible for the PDF versions of the documents to be easily created by the authors before ingest or by the archive upon acquisition. Also, the increase in the number of non-Adobe PDF tools and PDF files had perhaps assuaged some of the earlier concern about the proprietary nature of Adobe products. (However, note that it was this very increase in non-Adobe software to create and read PDF that led the information standards community to begin the PDF/A-1 initiative. See section 4.3 below.)

For many organizations, particularly in the national library community, PDF was viewed as a beneficial but supplementary version to be submitted along with XML. In the case of PubMed Central, PDF supplemented the SGML/XML format by serving as an authoritative copy against which the SGML/XML could be validated before its inclusion in the PubMed Central archive. PDF also provided a guide for future rendering of the material by maintaining the look and feel of the original text object. The Royal Technology Library of Sweden kept the native format, generally Word or TeX/LaTeX, and then created a PDF version. However, the Library did not consider PDF to be a preservation format because of its proprietary nature.

The National Center for Biotechnology Information at the NLM developed the Archiving and Interchange DTD Suite. The purpose was to "…preserve the intellectual content of journals independent of the form in which that content was originally delivered." (NCBI, 2006) The suite provides a series of modules using XML. According to the web site, "the Archiving and Interchange DTD may be used as is, or the Suite can be used to construct DTDs for authoring and archiving journal articles as well as DTDs for transferring journal articles from publishers to archives and between archives." The Journal Archiving component of the suite is used by publishers to submit

content to PubMed Central. Note that the goal is to store the **content** in an independent form. This differs significantly from the goal of PDF, which is to store the **layout** and render the layout across platforms.

In 2004, archives reported receiving a variety of bitmapped image formats including JPEG and GIF. However, many institutions converted these formats to TIFF to preserve the best image in the most standardized format that is not subject to loss or compression. For example, NLM's Profiles in Science creates collections of important papers, videos, audios, and even e-mails from noteworthy scientists in biomedicine, particularly Nobel Laureates. The original paper document is retained, whether electronic or paper. The staff creates the highest quality TIFF possible and any browser formats are created from the TIFF. However, by retaining any original paper documents, the door is open for creating better access formats in the future by reprocessing the original.

PubMed Central requires original digital image files for all figures, as well as tables and equations that are constructed as images and are not encoded in the SGML or XML. PubMed Central requests lossless compression TIFFs or EPS (Encapsulated Postscript); JPEG and GIF may be sent if they are the only formats available. PubMed Central is anxious to receive the best quality image available. PubMed Central converted the TIFFs to JPEGs and GIFs for display on the web.

### *4.3 The Advent of PDF/A-1*

The preservation format issue has been raised anew by the advent of PDF/A-1. Several organizations, including the Association for Information and Image Management (AIIM) and NPES decided to address the preservation issues that were arising with the widespread use of PDF. The Administrative Office of the U.S. Courts was a driving force in forming a U.S. Committee to initiate an ISO standard based on PDF. A major goal was "to address the issue that large bodies of official documents and important information are maintained in PDF, but that PDF is not suitable as an archival format." (Arms & Fleischhauer, 2006**)** Once the effort was established, The National Archives and Records Administration (NARA) joined the discussions to represent the archival community in the standards process, influence the development process so that electronic records in PDF/A-1 format can be preserved by NARA over the long term, and to obtain information used in developing NARA guidance for transferring permanent records in PDF. (Redman, 2006)

The primary reason for developing an archival version of PDF was to address the variation in the file format caused by multiple vendors implementing the open PDF specification in different ways. Secondarily, the aim was to eliminate PDF features that can complicate preservation. The feature-rich nature of PDF can create difficulties in preserving PDF information over the long term. For example, PDF documents are not necessarily self-contained. Some PDF 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." (Sullivan, 2006a)

An early challenge was agreeing on the scope of the standard. There were many discussions regarding, for example, when PDF/A-1 would be applied in the document lifecycle and how to address compression restrictions. After many lengthy discussions, the group limited the standard to specify a file format. This left to its implementers "specific processes for converting paper or

electronic documents to the PDF/A-1 format; specific technical design, user interface, implementation, or operational details of rendering; specific physical methods of storing these documents such as media and storage conditions; and required computer hardware and/or operating systems." (Sullivan, 2006a) To address this implementation flexibility, the group emphasizes in the Introduction to the standard that PDF/A-1 does not stand alone. A "Best Practices statement" in Annex B details the capture and conversion processes that help ensure accurate replication of source data.

In 2005, PDF/A-1 was approved as an ISO Standard 19005-1: 2005, under TC 171. (It was issued and approved as PDF/A-1 to indicate that additional parts will be added based on future versions of PDF.) PDF/A-1 is a set of rules for what NOT to do in a PDF in order to have some chance of reading it later. It also specifies what is allowed and what is required in PDF/A-1 and how to implement those objects. This stricter definition is essential to understanding the appropriate uses for PDF/A-1 and the limitations of PDF/A-1as a default format for archiving electronic documents in general.

More specifically, PDF/A-1 is a constrained form of Adobe PDF version 1.4 intended to be suitable for long-term preservation of page-oriented documents for which PDF is already being used in practice.

According to the PDF/A-1-1 FAQ, … "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 "flavor" of PDF that is more suitable for archival preservation than traditional PDF. PDF/A-1 aims to preserve the static visual appearance of electronic documents over time and also aims 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. The result is a file format, based on PDF 1.4 that is more suitable for long term preservation. PDF/A-1 files will be more self-contained, self-describing, and more device-independent than traditional PDF 1.4 files." (AIIM, 2006)

According to the LC-NDIIPP assessment, PDF/A-1 attempts to maximize device independence, self-containment, and self-documentation, which are all factors considered beneficial to the sustainability of a format, and, therefore, desirable for preservation purposes. However, PDF/A-1 would not be appropriate for all materials that can use the PDF format. Audio and video content, javascript and executable file launches, as well as encryption are prohibited. All fonts must be legally embeddable for unlimited universal rendering, and colorspaces must be specified in a device-independent way. Standards-based metadata is mandated. (Arms & Fleischhauer, 2006)

As with many standards, it is important to consider their scope and purpose. Often, standards are misunderstood because they are stretched to serve purposes for which they were never intended. The writings of Sullivan and Fanning and the AIIM FAQ for PDF/A-1 provide insights into the history, purpose, and scope of PDF/A-1. As Susan Sullivan, representative from NARA to the PDF/A development committee states, "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." (Sullivan, 2006a)

### *4.4 The Current Situation*

Many of the preservation approaches in place in 2004 continue to be used today. In practice, most organizations use a variety of formats as the basis for their operational systems. However, research and industry groups have continued to address the problem of digital preservation formats. Specific enterprises concerned with the archiving and preservation of scientific and technical information have made pragmatic decisions and, in some cases, established preferences for formats.

The National Center for Biotechnology Information at the NLM has extended the Archiving and Interchange Suite to include electronic books and online documentation. An increasing number of primary and secondary publishers, including JSTOR and CSIRO, have based their own efforts on the DTD Suite provided by NCBI/NLM. The Library of Congress and the British Library have voiced support for NLM's DTD. (Library of Congress, Digital Preservation, 2006)

The LC has suggested preferences for different types of information including text, indicating that if information is available in XML or some other structural mark-up, this format is preferred. PDF/A-1 is also an acceptable format as is PDF, if that is available. "PDF/A is suggested as a preferred format for page-oriented textual (or primarily textual) documents when layout and visual characteristics are more significant than logical structure. More proprietary formats, such as Microsoft Word binary format, are not generally suitable for LC collections. The preferences are based on an analysis in 2006 of the sustainability of various formats documented on the Sustainability of Formats Web site." (Arms & Fleischhauer, 2006)

The US Government Printing Office has an initiative to digitize its legacy collection and is advocating digitization as a legitimate preservation strategy. In its role as coordinator, in partnership with the Federal Depository Library Program and others, the GPO is working to establish standards and best practices for this digitization. In February 2006, GPO released version 3.3 of its specifications for converted content, which calls for the use of TIFF. (GPO 2006a) This version aligns the previous specifications with the development of the new GPO's Future Digital System (FDSys) production system. (GPO 2006b)

Similarly, GPO is concerned about providing continued access to material born digitally. While GPO is interested in standards and promoting appropriate practices for digital materials, the new infrastructure initiative, FDSys, is similar to the ERA in that it has committed to a plug and play approach, which is based on a variety of formats that, at any given point in time, may be supported to different degrees. As technologies change, new formats will be added, and as technologies and practices become available for migrating and preserving these formats, these solutions will be added to the toolbox.

In March 2003, NARA issued its guidance for the deposit of PDF documents. (NARA, 2003a) By that time, the number of agencies using PDF had grown, PDF had begun to have a history of backward compatibility, and NARA was deeply focused on issues surrounding the deposit of electronic records. However, NARA has not endorsed PDF, PDF/A-1, or any other preservation format. However, if Federal agencies intend to use PDF/A-1 for their permanent records, they will need to meet the additional requirements in NARA's PDF transfer guidance. Essentially these additional requirements apply to scanned image quality and acceptable methods of embedding OCR'd text. NARA's Electronic Records Archive Project is aimed at developing an infrastructure that can deal with any format. (Cahoon, 2006) NARA views itself on the receiving end of a lifecycle that begins with agencies making decisions about formats based on their business needs, rather than on NARA's

acceptance of the format for permanent records. Therefore, there will always be formats that are easier to archive than others because they can be more reliably preserved and rendered over time, and cases where successful rendering of a format must wait for technology to become available before the next migration or transformation can occur.

To date, PDF/A-1 does not play a major role in agency preservation plans. It is likely that PDF/A-1 will find its early adopters among the records managers and archivists, particularly for administrative data. The widespread adoption of PDF/A-1 is closely linked to the availability of software. In addition, adopters must create the environment within which PDF/A-1 will be used, including the related policies and procedures. (See Section 6.0 below.)

## 5.0 Format Assessment

A major evaluation of the sustainability of digital formats has been conducted by LC to meet the needs of the Library. The evaluation is based on several LC-oriented factors: 1) consideration for the deposit of digital works under Copyright Law, 2) the acquisition needs of the Library, 3) the need for systems, automated tools, and workflow processes, and 4) the need to identify more specific technical requirements for formats that have already been accepted or are designated as preferable. However, this assessment and the framework for making decisions based on the assessment are widely applicable and have been recognized and used by LC's partners in the NDIIPP program.

### *5.1 Technical Factors*

Rather than recreate the technical assessment of the formats, the analysis for those formats of interest to this paper have been extracted from the Format Sustainability resource presented by the LC-NDIIPP. The assessments for the four formats analyzed in this paper are provided in Appendix A and summarized in the tables in the Executive Summary. In the NDIIPP analysis, each format of interest is analyzed against the factors for sustainability described in Section 2.0 above.

### *5.2 Quality and Functionality*

The LC-NDIIPP assessment highlights not only the technical factors mentioned earlier in this paper, but what it calls quality and functionality factors. Quality and functionality factors pertain to the ability of a format to represent the significant characteristics of a given content item required by current and future users. These factors will vary for particular genres or forms of expression for content. For example, significant characteristics of sound are different from those of still pictures, whether digital or not, and not all digital formats for images are appropriate for all genres of still pictures.

At a given point in time, the user's baseline requirements with regard to quality and functionality can be determined. The baseline requirements will, of course, change over time. However, users with advanced needs may have additional requirements that go beyond this common level. Depending on the purpose of the archive and the user groups it serves, these requirements may need to be addressed through special software or partnerships with other archives that preserve more of the characteristics of the object.

Quality and functionality can be viewed as the importance of content versus layout and presentation. To what degree does the archive want or need to insure flexibility in the re-use of content in the future? To what extent does the archive want or need to preserve the look and feel of the original?

### 5.2.1 Preserving Content for Re-use

There are two aspects to the re-use of content. The first is the ability to use parts of the content, re-bundling it with other content or using it separately. According to the NLM, an interesting thing happens when content is kept and made available over time: people find new uses for it or parts of it. (Beck, 2006) Similar to the recording industry where changes in media from LP to compact disc led to migration issues and obsolescence, text will become obsolete when constrained to a particular medium or proprietary format. Digitization of music led to reuse of content from the early samplings from multiple artists to today's "mashup", which has become a new genre that combines parts of different recorded songs to form a new one. Peer-to-peer technologies, such as Napster, resulted in the "unbundling" of music from albums to individual songs. Users collect and build their own music libraries song by song rather than buying a complete album. Similarly, articles from PubMed Central, which were formerly available as parts of issues, are available through the PMC archive individually. They can be searched and assembled into collections based on user's interests. This is possible because the content in PMC is maintained in XML.

The second aspect of re-use is being able to present the content in layouts and presentations different from its original. This can be as simple as changing the layout of the format on the page or as complicated as successfully rendering the content on a variety of hand-held or non-traditional devices, such as iPODs, PDAs and cell phones. This aspect of re-use may be especially important for organizations and disciplines such as medicine, materials science, environmental management or engineering, where "chunks" of content are increasingly useful in workflows that take place in non-traditional work environments or that benefit from the combination of pieces of information from a variety of sources.

XML is the format of choice for preservation of the content with a goal of re-use, re-purposing and re-presentation of content. XML separates the content from the presentation. Through tagging and the use of style sheets, the content can be rendered in its native form or presented in a number of different formats and on different devices. If the original is produced in XML with appropriate schema or DTDs and style sheets, the content can be preserved and the format can be preserved. Recreating the original requires bringing these two components back together again. However, the fact that the content is in a well-formed, well-documented XML format allows the content to "stand on its own".

### 5.2.2 Preserving Layout and Presentation

For some organizations and disciplines, retaining the look and feel of the original content may be as much or more important than the content itself. In this case, the significant characteristics have to do with the layout and presentation of the original. This can be particularly important for situations involving records management, evidence and citizens-right-to-know. In these cases, the issue may be less one of disseminating and integrating the information in the future and more about the content's container when a particular event took place or a decision was made. Similar to the records management principle of retaining a record in the context of other records, this aspect of preservation focuses on retaining the look and feel and assuring that the content can be rendered in exactly the

form and layout as the original. The original layout may have been a critical part of how users used, interpreted, and made decisions based on the information.

The ability to preserve the image of the content, layout and presentation is a hallmark of the TIFF image format. PDF can also preserve the presentation and layout with more functionality than TIFF. However, PDF suffers from the variant implementations across vendors.

The preservation of layout was central to the development of PDF/A-1; administrative records delivered in a variety of PDF versions and implementations needed to be made available. PDF/A-1 supports two conformance levels. Level A uses Tagged PDF and Unicode character maps to preserve the document's logic structure and content text stream in a natural reading order, supporting a higher level of document preservation services over time. Level B includes all requirements of ISO 19005-1 minimally necessary to preserve the visual appearance, allowing conformance to PDF/A-1 without requiring users to define structure or other descriptive information. (Sullivan, 2006a) While PDF/A-1 is a minimalist approach, potentially leaving unpreserved more advanced PDF features, a benefit to PDF/A-1 is that it encourages all stakeholders within an enterprise to consider the significant characteristics of the material and their impact on future preservation efforts when the material is created.

### *5.3 Striking a Balance*

In practice, the technical, quality and functionality factors surrounding preservation formats must be balanced. The LC-NDIIPP Format Sustainability Web site notes that sometimes these factors may directly compete with one another. "…Some formats adopted widely for delivery of content to end users are proprietary or apply lossy compression for transmission over low-bandwidth networks." For content of high cultural value and for which a special functionality has particular significance, the ability of a format to support that functionality may outweigh the sustainability factors. The acceptance of the format by the contributors or users of the archive and their ability to contribute to the archive may outweigh the benefits to be gained by a more transparent format. Setting standards that are not in line with the work habits of the contributors will often assure that the archive has no content to preserve. In these cases, adoption may be a more important factor than others. The choice of format for preservation that achieves the right balance may change over time, particularly as new formats are adopted for creation and use. The most appropriate format will also involve decisions related to policy, the retention of multiple formats to serve different purposes, the publication process for the material to be preserved, and the resources and costs associated with the various alternatives.

## 6.0 Preservation Formats as Part of the Archival Process

This paper has focused on the technical decisions surrounding preservation formats, but it is important to remember that selecting and implementing a preservation format alone do not ensure the longevity of digital information. Agencies must implement preservation formats along with policies and procedures to ensure the quality and integrity of the information. Annex B of ISO 19005 (the PDF/A-1 standard) acknowledges that "this part of ISO 19005 should be used as one component of an organization's electronic archival environment for long-term retention of documents. Successful implementation of this part of ISO 19005 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 - the records management standard]; any additional requirements and conditions necessary to ensure the persistence of electronic documents and their characteristics over time." (ISO 19005-1, 2005).

## 7.0 Conclusion

There are a number of factors that come into play when determining the most appropriate format for preservation. As with any attempt to make the most appropriate decision, the agency must clearly define the purpose and the requirements for preservation, and, therefore, the purpose and requirements for the preservation format. The appropriate answer will depend on the mission of the agency, the kind of information to be preserved, the uses to which the objects may be put in the future, the expectations of current and future users, and how far into the future the objects are intended to remain useful. Admittedly, there is no crystal ball, but pragmatic decisions require that factors be balanced.

In general, XML is the most open, least proprietary format. It also provides flexibility for re-use of the content. When the XML is well-documented and complemented by preservation of the look and feel through PDF or TIFF, it is the most appropriate for organizations looking to ensure re-use, re-presentation, and re-purposing of the content. For archives focused on making content available for use and reuse, the PDF format, even with the restrictions of PDF/A-1, is not flexible enough to build and maintain a reliable archive that can be migrated. The further development and widespread adoption of formats and tools for the creation of documents, such as the NLM Archiving Suite and the XML DTD for technical reports, will help to advance the use of XML from the beginning of the document's life cycle.

PDF is problematic as a long-term preservation format, because the PDF specification has been implemented over several versions and by many vendors. The rich features that have and will continue to be added to PDF based on market drivers will also result in PDF file formats that complicate the long-term preservation process.

For retaining the current look and feel, particularly in a records environment, or where PDF is the most likely format for incoming material, PDF/A-1 is an appropriate choice. PDF/A-1 should be considered if PDF is already the format of choice or the only format available, since it is intended to bridge versions of PDF over time and across vendors. The PDF/A-1 format is appropriate for content when the significant characteristics of the material are not lost when moving from PDF to PDF/A-1.

For situations focused on the look and feel, where image-only access without full-text searching will suffice, bitmapped images in the TIFF format provide a good solution. TIFF images are produced by most capture systems, and files in this format can be launched with many applications, including a variety of free plug-ins for browsers. Although bitmapped images have no native capability to store metadata or to render active hyperlinks, TIFF continues to be an appropriate output for digitization initiatives and is easily paired with other approaches, including PDF and XML.

In addition, as the descriptions of operational systems from 1999 to the present day show, many organizations have chosen to store multiple formats. This allows the use of one format to preserve the content for re-use and another to preserve the layout. A multi-format approach is also likely to support migration to more robust formats in the future.

Along with policies and procedures about content and workflow, the decision about the preservation format(s) is an important component of any digital preservation plan. The decision regarding the most appropriate preservation format must be made within a framework that balances cost, functionality, quality, and sustainability.

# References

AIIM. (2006). "Frequently Asked Questions (FAQs): ISO 19005-1: 2005". July 10, 2006. [Online]. Available: http://www.aiim.org/documents/standards/19005-1_FAQ.pdf [July 31, 2006]

Arms, C. & C. Fleischhauer. (2006). "Sustainability of Digital Formats: Planning for the Library of Congress Collections." [Online]. Available: http://www.digitalpreservation.gov/formats/index.shtml [July 31, 2006]

Arms, C. & C. Fleischhauer. (2005). "Digital Formats: Factors for Sustainability, Functionality and Quality." Paper presented at the 2005 I&ST Archiving Conference. Washington DC, April 2005. [Online.] Available: http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf [July 31, 2006]

Beck, J. (2003). "PubMed Central & the NLM DTDs." Presented at the ASIS&T DASER Summit held in Cambridge, MA, 21-23 November 2003. [Online]. Available: http://www.asis.org/Chapters/neasis/daser/Jeff_Beck_presentation.ppt

Beck, J. (2006). "NCBI comments on PubMed Central dated July 31, 2006."

Boyce, P. (1997). "Costs, Archiving, and the Publishing Process in Electronic STM Journals." *Against the Grain*, v. 9 #5,p. 86, Nov 1997. [Online]. Available: www.aas.org/~pboyce/epubs/atg98a-2.html [July 31, 2006]

Cahoon, L. R. (2006). "Building the Archives of the Future: The National Archives & Records Administration's Electronic Records Archives Initiative -- Architecture and Plans." Presentation at CENDI Meeting, February 9, 2006.

Carroll, B. & G. Hodge. (1999). "Digital Electronic Archiving: The State of the Art and the State of the Practice." Published by ICSTI and CENDI. April 1999. [Online]. Not Available; updated 2004. See Hodge, G. & E. Frangakis, in this List of References for updated document.

CENDI. (2007). "Model Statement on the Use of Digitization as a Preservation Format." (2007). [Online]. Available: http://www.cendi.gov/publications/model_digi_stmnt.pdf [March 2007]

Davis, R. (2006). Personal communication.

Fanning, Betsy. (2005). "PDF/A-1rchive An Update." *AIIM E-doc Magazine*. May/June 2005. [Online.] Available: http://www.edocmagazine.com/print.asp?ID=30118 [October 5, 2006]

GPO (US Government Printing Office). (2006a). FDsys Operational Specification for Converted Content (Version 3.3) (Feb. 2006). Digitization Specifications and Operating Procedures for Archiving Materials: Creation of Preservation Master Files For the following content types – Textual, Graphic Illustrations / Artwork, Originals, and Photographs. [Online.] Available: http://www.gpoaccess.gov/legacy/ [October 5, 2006]

GPO (US Government Printing Office). (2006b). "FDsys Status." [Online.] Available: http://www.gpo.gov/projects/fdsys.htm [October 5, 2006]

Harvard University Library. "Global Registry of Digital Formats." (2006). [Online]. Available: http://hul.harvard.edu/formatregistry/ [July 31, 2006]

Hodge, G. & E. Frangakis. (2004). "Digital Preservation and Permanent Access to Scientific Information: The State of the Practice." Published by CENDI. Revised April 2004. [Online.]. Available: http://www.cendi.gov/publications/04-3dig_preserv.pdf [July 31, 2006]

ISO. (2005). "ISO 19005-1, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1-1)"

Kyong-Ho, L., O. Slattery, R. Lu, X. Tang & V. McCrary. (2002). "The State of the Art and Practice in Digital Preservation." Journal of Research of the National Institute of Standards and Technology. Vol. 107, No. 1, January-February 2002, p. 93-106. [Online]. Available: http://nvl.nist.gov/pub/nistpubs/jres/107/1/j71lee.pdf

LC (Library of Congress), Digital Preservation. "Library of Congress, British Library to Support Common Archiving Standard for Electronic Journals." Press release. April 19, 2006. [Online.] Available: http://www.digitalpreservation.gov/news/pr_041706_2.html [October 5, 2006]

NARA (National Archives and Records Administration). (2003a). NARA Transfer Guidance for Records in PDF. March 31, 2003. [Online.] Available: http://www.archives.gov/records_management/policy_and_guidance/nwm11_2003.html [October 5, 2006]

NARA (National Archives and Records Administration). (2003b). Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records. RECORDS IN PORTABLE DOCUMENT FORMAT (PDF). March 31, 2003. [Online.] Available: http://www.archives.gov/records-mgmt/initiatives/erm-guidance.html [October 5, 2006]

NCBI (National Center for Biotechnology Information), National Library of Medicine. "Archiving and Interchange DTD." [Online.] Available: http://dtd.nlm.nih.gov/ [July 31, 2006]

NDIIPP (National Digital Information Infrastructure and Preservation Program). "Digital Preservation" (homepage). [Online.] Available : http://www.digitalpreservation.gov/ [July 31, 2006]

Redman, M. (2006). "NARA comments on PDF/A-1 dated October 13, 2006."

Sullivan, S. (2006a). "PDF/A-1: worldwide collaboration to preserve electronic documents." *ISO Focus.* March 2006. [Online.] Available: http://www.aiim.org/documents/standards/PDF-A-ISO-Focus.pdf [October 5, 2006]

Sullivan, S. (2006b). "An archival records management perspective on PDF/A-1." *Records Management Journal,* Vol. 16, Iss. 1, pp. 51-56. [Online by subscription only.] Available:

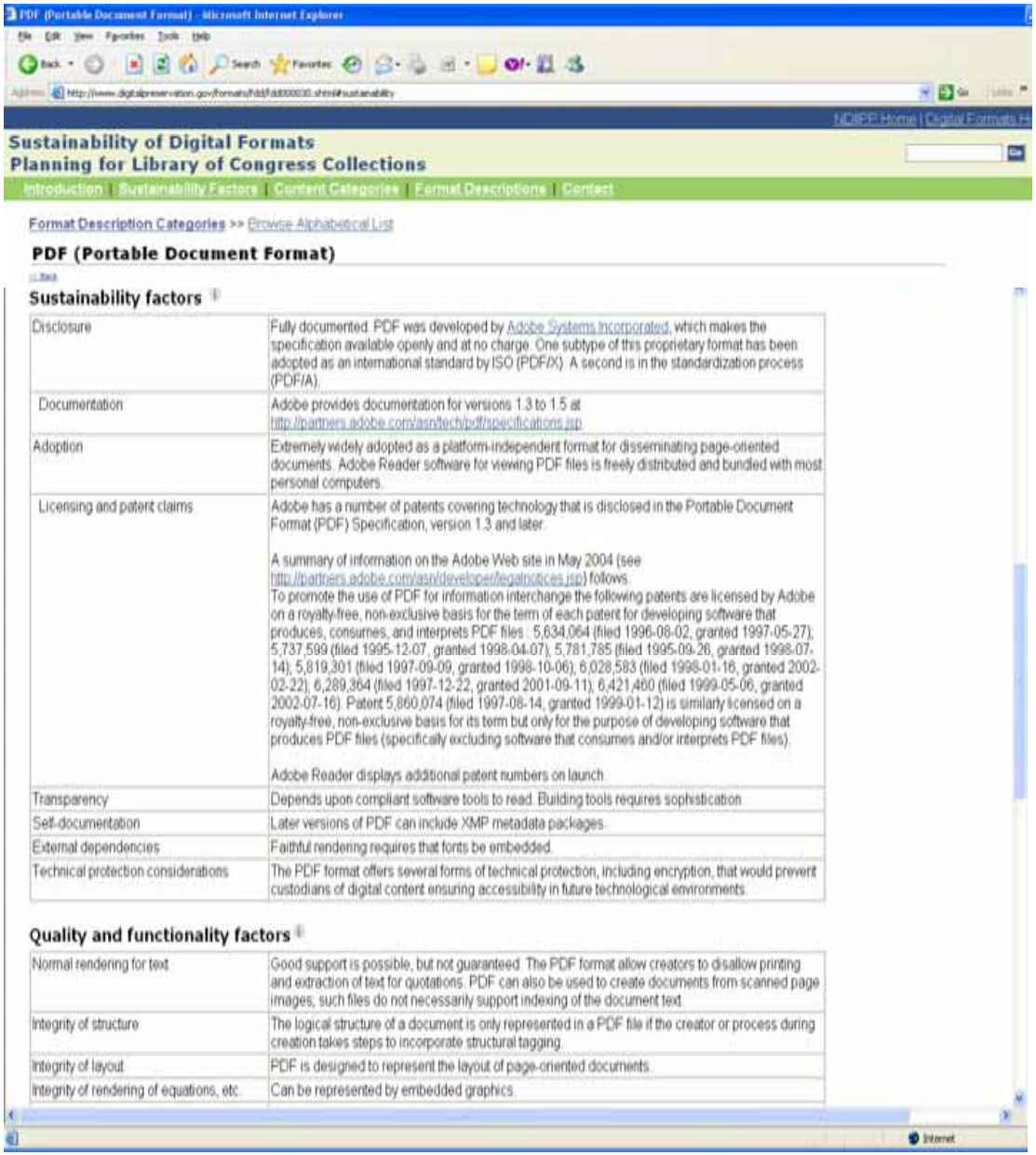
http://www.emeraldinsight.com/Insight/viewContainer.do?containerType=Issue&containerID=23699 [October 5, 2006]

# Appendix A

## Technical Assessment of Preservation Formats
## National Digital Information Infrastructure for Preservation Program

http://www.digitalpreservation.gov/form0.ats/fdd/fdd000125.shtml

**PDF (Portable Document Format)** **http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml**



Sustainability of Digital Formats
Planning for Library of Congress Collections

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

Format Description Categories >> Browse Alphabetical List

## PDF (Portable Document Format)

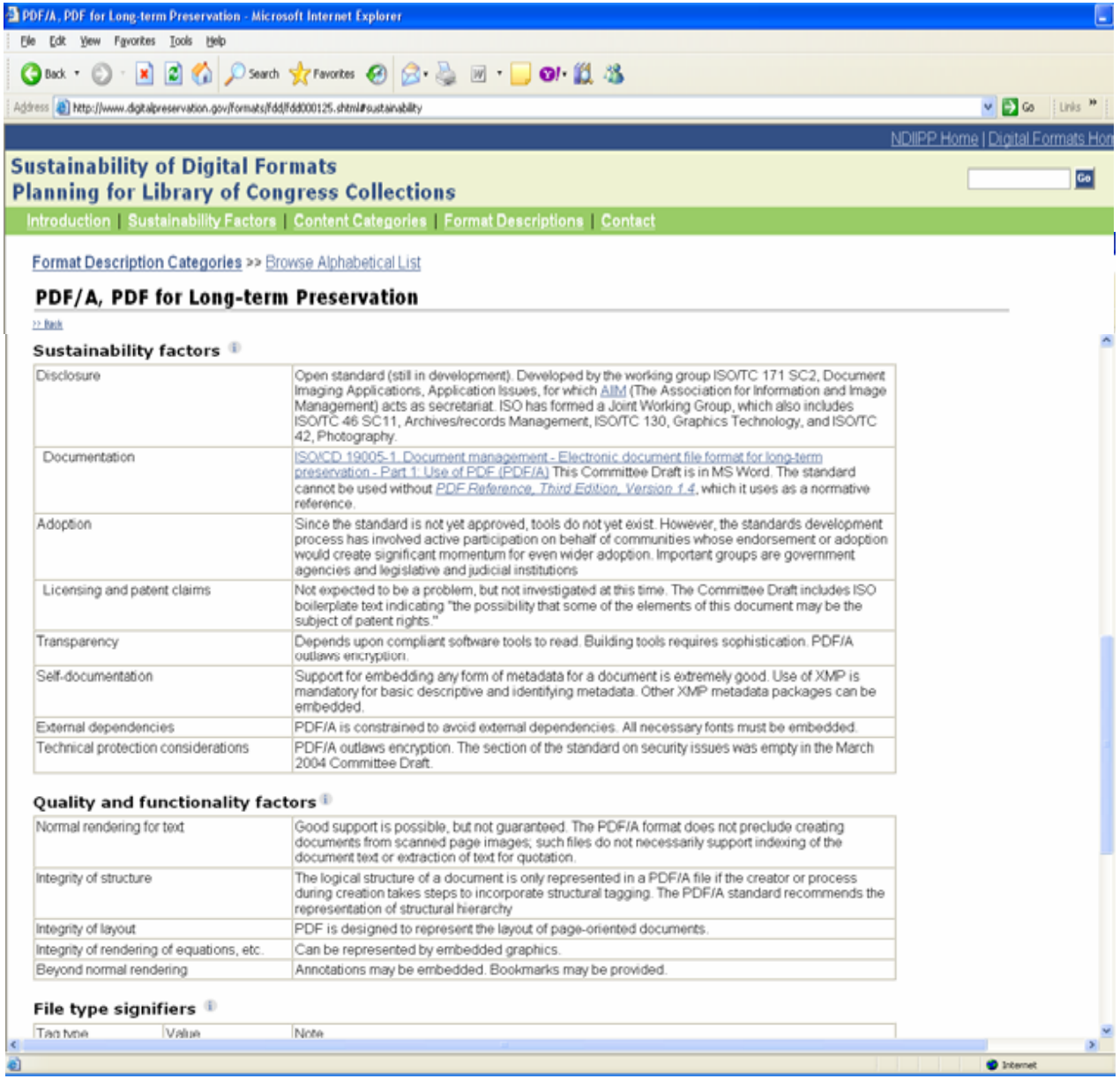
### Sustainability factors

| | |
|---|---|
| Disclosure | Fully documented. PDF was developed by Adobe Systems Incorporated, which makes the specification available openly and at no charge. One subtype of this proprietary format has been adopted as an international standard by ISO (PDF/X). A second is in the standardization process (PDF/A). |
| Documentation | Adobe provides documentation for versions 1.3 to 1.5 at http://partners.adobe.com/asn/tech/pdf/specifications.jsp. |
| Adoption | Extremely widely adopted as a platform-independent format for disseminating page-oriented documents. Adobe Reader software for viewing PDF files is freely distributed and bundled with most personal computers. |
| Licensing and patent claims | Adobe has a number of patents covering technology that is disclosed in the Portable Document Format (PDF) Specification, version 1.3 and later.<br><br>A summary of information on the Adobe Web site in May 2004 (see http://partners.adobe.com/asn/developer/legalnotices.jsp) follows:<br>To promote the use of PDF for information interchange the following patents are licensed by Adobe on a royalty-free, non-exclusive basis for the term of each patent for developing software that produces, consumes, and interprets PDF files : 5,634,064 (filed 1996-08-02, granted 1997-05-27); 5,737,599 (filed 1995-12-07, granted 1998-04-07); 5,781,785 (filed 1995-09-26, granted 1998-07-14); 5,819,301 (filed 1997-09-09, granted 1998-10-06); 6,028,583 (filed 1998-01-16, granted 2002-02-22); 6,289,364 (filed 1997-12-22, granted 2001-09-11); 6,421,460 (filed 1999-05-06, granted 2002-07-16). Patent 5,860,074 (filed 1997-08-14, granted 1999-01-12) is similarly licensed on a royalty-free, non-exclusive basis for its term but only for the purpose of developing software that produces PDF files (specifically excluding software that consumes and/or interprets PDF files).<br><br>Adobe Reader displays additional patent numbers on launch. |
| Transparency | Depends upon compliant software tools to read. Building tools requires sophistication. |
| Self-documentation | Later versions of PDF can include XMP metadata packages. |
| External dependencies | Faithful rendering requires that fonts be embedded. |
| Technical protection considerations | The PDF format offers several forms of technical protection, including encryption, that would prevent custodians of digital content ensuring accessibility in future technological environments. |

### Quality and functionality factors

| | |
|---|---|
| Normal rendering for text | Good support is possible, but not guaranteed. The PDF format allow creators to disallow printing and extraction of text for quotations. PDF can also be used to create documents from scanned page images; such files do not necessarily support indexing of the document text. |
| Integrity of structure | The logical structure of a document is only represented in a PDF file if the creator or process during creation takes steps to incorporate structural tagging. |
| Integrity of layout | PDF is designed to represent the layout of page-oriented documents. |
| Integrity of rendering of equations, etc. | Can be represented by embedded graphics. |

## PDF/A-1, PDF for Long-term Preservation http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml



PDF/A, PDF for Long-term Preservation - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml#sustainability

NDIIPP Home | Digital Formats Hom

**Sustainability of Digital Formats**
**Planning for Library of Congress Collections**

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

Format Description Categories >> Browse Alphabetical List

### PDF/A, PDF for Long-term Preservation

<< Back

#### Sustainability factors

| | |
|---|---|
| Disclosure | Open standard (still in development). Developed by the working group ISO/TC 171 SC2, Document Imaging Applications, Application Issues, for which AIIM (The Association for Information and Image Management) acts as secretariat. ISO has formed a Joint Working Group, which also includes ISO/TC 46 SC11, Archives/records Management, ISO/TC 130, Graphics Technology, and ISO/TC 42, Photography. |
| Documentation | ISO/CD 19005-1, Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF (PDF/A) This Committee Draft is in MS Word. The standard cannot be used without PDF Reference, Third Edition, Version 1.4, which it uses as a normative reference. |
| Adoption | Since the standard is not yet approved, tools do not yet exist. However, the standards development process has involved active participation on behalf of communities whose endorsement or adoption would create significant momentum for even wider adoption. Important groups are government agencies and legislative and judicial institutions |
| Licensing and patent claims | Not expected to be a problem, but not investigated at this time. The Committee Draft includes ISO boilerplate text indicating "the possibility that some of the elements of this document may be the subject of patent rights." |
| Transparency | Depends upon compliant software tools to read. Building tools requires sophistication. PDF/A outlaws encryption. |
| Self-documentation | Support for embedding any form of metadata for a document is extremely good. Use of XMP is mandatory for basic descriptive and identifying metadata. Other XMP metadata packages can be embedded. |
| External dependencies | PDF/A is constrained to avoid external dependencies. All necessary fonts must be embedded. |
| Technical protection considerations | PDF/A outlaws encryption. The section of the standard on security issues was empty in the March 2004 Committee Draft. |

#### Quality and functionality factors

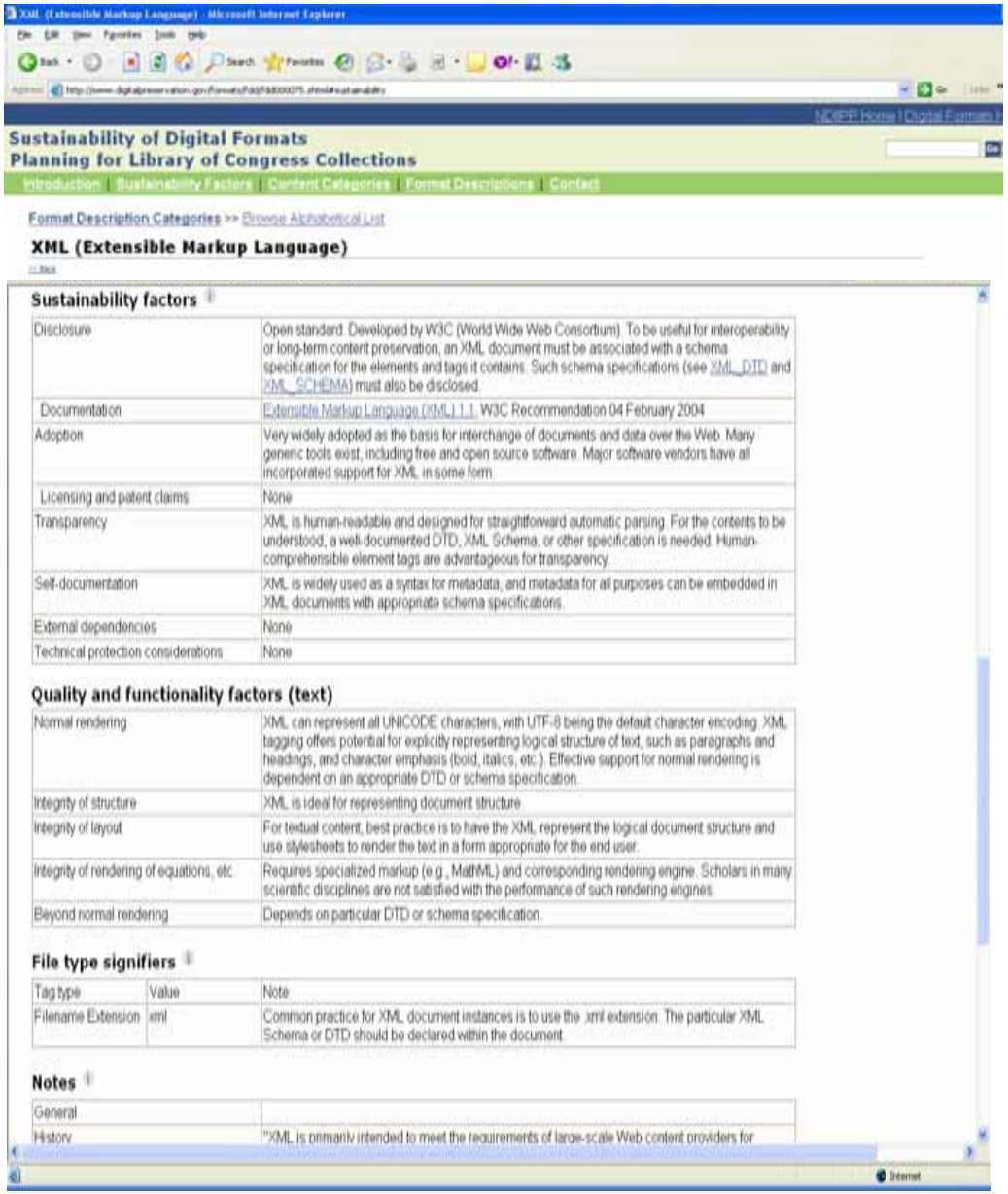
| | |
|---|---|
| Normal rendering for text | Good support is possible, but not guaranteed. The PDF/A format does not preclude creating documents from scanned page images; such files do not necessarily support indexing of the document text or extraction of text for quotation. |
| Integrity of structure | The logical structure of a document is only represented in a PDF/A file if the creator or process during creation takes steps to incorporate structural tagging. The PDF/A standard recommends the representation of structural hierarchy |
| Integrity of layout | PDF is designed to represent the layout of page-oriented documents. |
| Integrity of rendering of equations, etc. | Can be represented by embedded graphics. |
| Beyond normal rendering | Annotations may be embedded. Bookmarks may be provided. |

#### File type signifiers

| Tag type | Value | Note |
|---|---|---|

Internet

## XML (Extensible Markup Language) http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml



Sustainability of Digital Formats
Planning for Library of Congress Collections

Introduction | Sustainability Factors | Content Categories | Format Descriptions | Contact

Format Description Categories >> Browse Alphabetical List

**XML (Extensible Markup Language)**

**Sustainability factors**

| | |
|---|---|
| Disclosure | Open standard. Developed by W3C (World Wide Web Consortium). To be useful for interoperability or long-term content preservation, an XML document must be associated with a schema specification for the elements and tags it contains. Such schema specifications (see XML_DTD and XML_SCHEMA) must also be disclosed. |
| Documentation | Extensible Markup Language (XML) 1.1. W3C Recommendation 04 February 2004 |
| Adoption | Very widely adopted as the basis for interchange of documents and data over the Web. Many generic tools exist, including free and open source software. Major software vendors have all incorporated support for XML in some form. |
| Licensing and patent claims | None |
| Transparency | XML is human-readable and designed for straightforward automatic parsing. For the contents to be understood, a well-documented DTD, XML Schema, or other specification is needed. Human-comprehensible element tags are advantageous for transparency. |
| Self-documentation | XML is widely used as a syntax for metadata, and metadata for all purposes can be embedded in XML documents with appropriate schema specifications. |
| External dependencies | None |
| Technical protection considerations | None |

**Quality and functionality factors (text)**

| | |
|---|---|
| Normal rendering | XML can represent all UNICODE characters, with UTF-8 being the default character encoding. XML tagging offers potential for explicitly representing logical structure of text, such as paragraphs and headings, and character emphasis (bold, italics, etc.). Effective support for normal rendering is dependent on an appropriate DTD or schema specification. |
| Integrity of structure | XML is ideal for representing document structure. |
| Integrity of layout | For textual content, best practice is to have the XML represent the logical document structure and use stylesheets to render the text in a form appropriate for the end user. |
| Integrity of rendering of equations, etc. | Requires specialized markup (e.g., MathML) and corresponding rendering engine. Scholars in many scientific disciplines are not satisfied with the performance of such rendering engines. |
| Beyond normal rendering | Depends on particular DTD or schema specification. |

**File type signifiers**

| Tag type | Value | Note |
|---|---|---|
| Filename Extension | xml | Common practice for XML document instances is to use the .xml extension. The particular XML Schema or DTD should be declared within the document. |

**Notes**

| | |
|---|---|
| General | |
| History | "XML is primarily intended to meet the requirements of large-scale Web content providers for |

## TIFF Bitmap with Group 4 Compression http://www.digitalpreservation.gov/formats/fdd/fdd000024.shtml