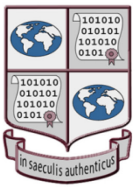

InterPARES 3 Project

International Research on Permanent Authentic Records in Electronic Systems

TEAM Korea

Title: General Study 20 – Examining the Criteria for Open Standard File Formats: General Study Report

Status:	Final (public)
Version:	4.0
Dated Submitted:	February 2012
Last Revised:	February 2016
Author:	The InterPARES 3 Project, TEAM Korea
Writer(s):	Eun G. Park School of Information Studies, McGill University, Montréal, Québec
Project Unit:	Research
URL:	http://www.interpares.org/display_file.cfm?doc=ip3_korea_gs20_final_report.pdf

Document Control

Version history			
Version	Date	By	Version notes
3.0	2012-02-13	E. Park	First compilation of draft final report.
3.1	2012-02-16	A. Allen	Minor formatting and grammatical changes.
4.0	2016-02-17	R. Preston	Content and copy edits for public version.

Table of Contents

General Study Report

1. Overview

Long-term preservation of electronic documents is one of the main challenges of modern archives. As Becker et al. mentioned, the ever-growing complexity and heterogeneity of digital file formats have evolved technically and as these changes bring challenges to the preservation of digital information (2008) we are faced with a multiplicity of file formats. Most file formats that currently exist are proprietary and dependent on various operating systems, hardware and software combinations. Archivists need to choose proper file formats by using a few methods to preserve electronic records. Many researchers have proposed several kinds of their own criteria. This research intends to examine the various criteria described in the literature in the major families to identify the basic requirements of appropriate file formats for long-term preservation and to access and help select the best file formats. We hope that an examination of the appropriateness of various standards will help determine the basic requirements for choosing file formats for the long-term preservation of electronic documents.

2. Comparison of Major File Formats

Standard file formats owe their status to (official) initiatives for standardizing or to their wide use (DAVID project, 2003). Some of the standard file formats that are accepted and recommended by national or international standards organizations include the following: TIFF/IT (ISO 12639: 2004), PNG (ISO/IEC 15948: 2004), JPEG 2000 (ISO/IEC 15444: 2003, 2004, 2005, 2007, 2008), PDF/A-1 (ISO Standard 19005-1 under TC 171 (ISO, 2005) *Document Management – Electronic Document File Format for Long-Term Preservation*), Open Office XML, Open Document Format, etc. Among these formats, four digital formats – Tagged Image File Format (TIFF), portable document format (PDF), PDF/A and XML are most frequently evaluated in various settings and contexts.

2.1 TIFF

TIFF (which originally stood for Tag or Tagged Image File Format) was developed in the mid-1980s by the desktop publishing industry. Aldus Corporation, which also developed PageMaker, developed TIFF to attempt to standardize file exchange of images between content

producers. The company was later acquired by Adobe (which now owns the rights to both PageMaker and TIFF). Unlike most other image file formats, TIFF is a 'lossless' format, which means that no information is lost when creating the file from the original file. (By contrast, JPEG is an example of a 'lossy' image file format, as it leads to some loss of information because of its use of compression.) TIFF now forms the basis of several specifications, including TIFF/EP (ISO 12234-2) used in electronic photography and TIFF/IT (ISO 12639) used for image technology (Wikipedia, 2010). The main purpose of TIFF files is to document the physical appearance of a document in an image format. At present, TIFF files cannot be text indexed, but software developments and the use of optical character recognition (OCR) will make this option possible in the near future by appending the text as metadata to the TIFF image (Microsoft, 2010).

2.2 PDF and PDF/A

PDF (Portable Document Format) was developed in the early 1990s by Adobe Systems as a proprietary format and launched in 1993 (Adobe, 2010). The aim of the PDF file format was to encapsulate all the information necessary for full and proper display of electronic documents in a way that did not require a specific piece of expensive and/or proprietary software. Although the original versions of PDF did in fact require a proprietary piece of software, Adobe Reader is now freely available on the Web to view PDF documents, while Adobe Acrobat is available for a fee and used to create PDF documents. Since 2008, PDF has been an open standard published as ISO 32000. Various subsets of PDF are in use today, including PDF/A for archival purposes, PDF/E for the exchange of engineering drawings and PDF/X used by the printing and graphic arts industry. One of the main advantages of PDF and PDF/A files is that they are generally considered text documents, making them searchable, indexable and retrievable by information systems while preserving the physical layout of the document.

Many studies suggest that PDF is not a good long-term preservation format based on their own sets of evaluation criteria, including open availability, interoperability, processibility, and so on (Barnes, 2006; CENDI, 2007; Hodge & Anderson, 2007; Potter, 2006). Mainly because PDF is a proprietary file format owned by Adobe, different versions of PDF are not completely compatible with one another. Because PDF documents may depend on system fonts and extra features of a PDF, including compression, encryption, and so on, to be viewed, this may hinder the accessibility of PDF files (Barnes, 2006; CENDI, 2007; Potter, 2006; Hodge & Anderson,

2007). More importantly, the content of a PDF file might not be in natural reading order, is not human readable without a proper PDF reader, and is in a binary file format, which implies that a bit corruption may cause catastrophic failure (Barnes, 2006). Potter (2006) also points out that the real weakness of PDF is its inability to be backed out into a usable format. Because of the widespread use of PDF, efforts have been put towards making PDF more amenable to long-term preservation. PDF/A-1 based on PDF 1.4 was approved as the International Organization for Standardization (ISO) Standard 19005-1 in 2005. "PDF/A-1 aims to preserve the static visual appearance of an electronic document over time and also aims to support future access and future migration needs by providing a framework for embedding metadata about electronic documents and defining the logical structure and semantic properties of electronic documents" (Sullivan, 2006, p. 52). Rog and Wijk (2007) apply their quantitative method to assess the Microsoft Word 97-2003 document format and PDF/A-1. PDF/A-1 scores 89.01% in comparison to 21.83% scored by Microsoft Word. They note that it is easy to reach a consensus on criteria, but not everyone agrees on the importance of each criterion.

2.3 XML

XML (eXtensible Markup Language) was developed under the direction of the World Wide Web Consortium (W3C). The SGML Editorial Review Board (now the XML Working Group) determined a few criteria that would be key to the launch, adoption and success of XML as an open standard, including that it would be compatible with SGML, human-legible and clear to understand, formal, and that documents would be easy to create (W3C, 2008). The W3C officially recommended XML Version 1.0 in 2008. Numerous subsets of XML have appeared since then and are used in various fields. RDF and Dublin Core can be expressed in XML. XML is purely open standard, as defined below. It is documented and maintained by the W3C, in collaboration with various industry stakeholders, and changes to the standard are governed by a general consensus approach (World Wide Web Consortium, 2009).

Sullivan (2006, p. 10) explained what the main purpose of long-term preservation is. When specifically considering PDF/A, she argued that: "The future use of, and access to, these objects [referring to PDF files] 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." This applies not only to PDF/A, but also to all long-term

preservation file formats. Issues of maintaining visual appearance, high-order properties (i.e., the document's structure) and metadata are part of the daily routine of archivists and records managers. Although the archivist is concerned with long-term preservation in the name of cultural and societal preservation, the records manager is concerned with long-term preservation in the name of administrative, financial, or economic requirements. However, in both cases, they must maintain access to these documents and, in some cases, provide for mechanisms for the repurposing of the information contained within.

In addition, the Office Open XML (OOXML or OpenXML) specification is an open standard file format recognized by ISO and IEC as an International Standard (ISO/IEC 29500). Meanwhile, Open Document Format (ODF) has brought about attention with support from the fast-growing open source community. The format became an Organization for the Advancement of Structured Information Standards (OASIS) standard and was approved as ISO/IEC 26300:2006. Microsoft also published its specification of Office Open XML in 2005, which was accepted as ISO/IEC DIS 29500:2008. OOXML, ODF and other formats are assessed in numerous studies. Barnes (2006) compares Microsoft Word (.doc), ODF, PDF, Rich Text Format (RTF), and two specific XML formats, DocBook XML and Text Encoding Initiative (TEI), based on four criteria: content-level, not presentation-level description; ample comment space, open availability, and interpretability, as proposed by Lesk (1995).

3. Methodology

The principal methodology used for this general study was content analysis. Graduate Research Assistants (hereafter GRAs) conducted a comprehensive review of existing studies that deal with file formats with their own criteria, including journal articles, reports, and grey literature. This was to provide an overview of necessary characteristics of and criteria for file formats. A table was produced to compile the criteria for assessing file formats.

4. Findings

Based on the findings, we grouped various criteria into four major families by their commonality: autonomy, interoperability, authenticity and functionality. The findings are provided in the following table.

Table 1. File Formats for Long-term Preservation

* Family 1 = Autonomy Family, Family 2 = Interoperability Family, Family 3 = Authenticity Family and Family 4 = Functionality Family.

Family*	Criteria	Definition/Notes	Referred by	Other Formats
1	Disclosure	Publically available authoritative specification.	Abrams et al. (2005)	PDF/A (Yes) MS Word (No)
1	Device Independencies	Can be reliably and consistently rendered without regard to the hardware/software platform.	Abrams et al. (2005)	PDF/A (Yes) TIFF (No)
1	Self-documenting	To contain its own description.	Abrams et al. (2005)	N/A
1	Self-contained	To contain all resources necessary for rendering.	Abrams et al. (2005)	N/A
1	Ample Comment Space	To allow rich metadata.	Barnes (2006)	N/A
1	Content-Level Description	Not presentation-level description. Structural markup, not formatting.	Barnes (2006)	PDF (No) DocBook (Yes) TEI (Yes) XHTML (Yes)
1	Open Availability	No proprietary formats.	Barnes (2006)	ODF (Yes) GIF (No) PDF (No) RTF (No) MS Word (No)
1	Open Standard	Formats for which the technical specification has been made available in the public domain.	Brown (2003)	JPEG (Yes) PDF (Limited) ASCII (Limited)
1	Metadata Support	Some file formats make provision for the inclusion of metadata.	Brown (2003)	TIFF (Yes) MS Word 2000 (Yes)

1	Stability	The format specification should be stable and not subject to constant or major changes over time. A new version of the format should also be backwards.	Brown (2003)	N/A
1	Disclosure	Existence of complete documentation.	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes)
1	Self- Documentation	Digital objects that contain basic descriptive, technical, and other administrative metadata.	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes)
1	Impact of Patents	Degree to which the ability of archival institutions to sustain content in a format will be inhibited by patents.	CENDI (2007) Hodge & Anderson (2007)	PDF (Not expected to be a problem) PDF/A (Yes) TIFF_G4 (No)
1	Raw I/O Efficiency	Formats that are organized for fast sequential access. Formats that aggregate many large objects in a single file can also be beneficial for this kind of application.	Folk & Barkstrom (2003)	N/A
1	Ease of Subsetting	The file format needs to support efficient extraction of irregularly-shaped subsets of array elements – and perhaps similarly shaped subsets from several arrays to put this requirement more carefully.	Folk & Barkstrom (2003)	N/A
1	Rigorous Definition	It is important that the format be defined in a sufficiently rigorous way so that readers can be written that correctly interpret the content of data files.	Folk & Barkstrom (2003)	N/A
1	Self-describing	Many different types of metadata are required to decipher the content of a file. Thus, the extent to which a file is self-describing is one measure of its suitability as an archive format.	Folk & Barkstrom (2003)	N/A
1	No Definite Term	Its self-describing tags identify what the content is all about.	Johnson (1999)	N/A

1	Content-level, not presentation- level, descriptions	Where possible the labeling of items should reflect their meaning, not their appearance.	Lesk (1995)	SGML (Yes)
1	Ample Comment Space	Items should be labeled, as far as possible, with enough information to serve for searching or cataloging.	Lesk (1995)	TIFF (Yes)
1	Open Availability	Any manufacturer or researcher should have the ability to use the standard, rather than having it under the control of only one company.	Lesk (1995)	Kodak PhotoCD (No) GIF (No)
1	No Definite Term	XML was discussed early on as a format for strong descriptive and administrative metadata and the complete content of the document.	Müller et al. (2003)	N/A
1	No Definite Term	Non-proprietary format.	Potter (2006)	PDF (Yes)
1	No Definite Term	Without consistent metadata, users can expect their searches to "hide desired records".	Potter (2006)	PDF (No)
1	No Definite Term	To avoid vendor-lock.	Potter (2006)	ODF (Yes)
1	Openness	The characteristics (standardization, restrictions on the interpretation of the file format, reader with freely available source) indicate the relative ease of accumulating knowledge about the file format structure.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Yes) MS Word (No)
1	Self- Documentation	The characteristics (metadata and technical description of format embedded) indicate the format possibilities concerning encapsulation of metadata.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Limited) MS Word (Limited)
1	Metadata Support	Not Defined	Sahu (2006)	N/A
1	Device Independencies	PDF/A requires device independent components so that static visual appearance can be reliably and consistently rendered and printed without regard to the hardware or software platform used.	Suillivan (2006)	PDF/A (Yes) PDF/X (Yes)

1	Self-describing Files	PDF/A requires Adobe extensible metadata platform (XMP) be used for embedding metadata in PDF files.	Suillivan (2006)	PDF/A (Adobe Extensible Metadata Platform Required)
1	Disclosure	PDF/A is based on an authoritative specification is publicly available. Anyone can use the PDF reference and XMP specification in conjunction with PDF/A to create applications that read, write, or process PDF/A files.	Suillivan (2006)	PDF/A (Yes)
1	Self- containment	Everything that is necessary to render or print a PDF/A file must be contained within the file.	Suillivan (2006)	PDF/A (Yes)
1	Accessibility	To prohibit encryption in the file trailer means that User IDs and/or Passwords are not needed to do anything with a PDF/A file.	Suillivan (2006)	PDF/A (Yes)
2	Adoption	Widespread use may be the best deterrent against preservation risk.	Abrams et al. (2005)	TIFF (Yes)
2	Interpretability	The formats should not binary. It should be possible for a human to read the data, and also for small errors in storage and transmission to remain localized.	Barnes (2006)	RTF (Yes) MS Word (No)
2	Ubiquity	The laws of supply and demand dictate that formats which are well established and in widespread use will tend to have broader and longer- lasting support from software suppliers than those which only have a niche market. Popular formats, which are supported by as wide of software as possible, are therefore to be preferred where possible.	Brown (2003)	N/A
2	Interoperability	The ability to exchange electronic records with other users and IT systems is frequently an important consideration. Formats which are supported by a wide range of software or are platform-independent are therefore highly desirable in many situations.	Brown (2003)	N/A

2	Processability	Certain types of data must retain their processability to have any reuse value, even though the requirements of authenticity demand that the archived version must not be altered through reprocessing. The requirement to maintain a processable version of the record must therefore be considered.	Brown (2003)	Conversion of a word-processed document into PDF format. (No)
2	Adoption	Degree to which the format is already in use.	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes)
2	Transparency	Degree to which the digital representation is open to direct analysis.	CENDI (2007) Hodge & Anderson (2007)	PDF (Limited) PDF/A (Limited) TIFF_G4 (Limited)
2	External-dependency	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.	CENDI (2007) Hodge & Anderson (2007)	PDF (Limited) PDF/A (No) TIFF_G4 (No)
2	Popularity	A format that is widely used is more likely to have either commercial or Open Source readers available.	Folk & Barkstrom (2003)	N/A
2	Availability of Readers	One way to maintain ease of data access is to ensure that readers are available for accessing archived data files.	Folk & Barkstrom (2003)	N/A
2	Ability to Embed Data Extraction Software in the files	The files come with read software embedded. Users get a self-extracting file that installs itself after downloading.	Folk & Barkstrom (2003)	N/A
2	Ease of Implementing Readers - Simplicity	If readers are not available for a particular file format, but the file format is simple, it may be easy to write readers from scratch.	Folk & Barkstrom (2003)	N/A
2	Ability to Name File Elements	To work with data based on manipulating the element names instead of binary offsets, or other references.	Folk & Barkstrom (2003)	N/A
2	Long-term Institutional Support	To ensure the long-term maintenance and support of a data format by placing responsibility for these operation on institutions, rather than individuals or projects.	Folk & Barkstrom (2003)	N/A

2	Suitability for a Variety of Storage Technologies	The format is not geared toward any particular technology.	Folk & Barkstrom (2003)	N/A
2	Stability	Compatibility between versions.	Folk & Barkstrom (2003)	N/A
2	Formal (BNF- or XML-like) Description of Format	In the situation where language die or communities that produced data disappear, archives may need to retain the ability to create new readers solely on the basis of formal descriptions of the file contents.	Folk & Barkstrom (2003)	N/A
2	Multi-language Implementation of Library Software	One defense against language obsolescence is to have multiple implementations of readers for a single format.	Folk & Barkstrom (2003)	N/A
2	Open Source Software or Equivalent	Data centres and archives need to move toward obtaining Open Source arrangements for all parts of the file format and associated libraries.	Folk & Barkstrom (2003)	N/A
2	No Definite Term	Data is easily repurposed via tags or translated to any medium.	Johnson (1999)	N/A
2	No Definite Term	Data types map easily among different applications, so it is very interoperable.	Johnson (1999)	PDF/A (No) PDF (No) TIFF (No)
2	No Definite Term	Creating, using and reusing tags is easy, making highly extensible.	Johnson (1999)	N/A
2	No Definite Term	To make transferring data easy.	Johnson (1999)	N/A
2	Interpretability	The standard should be written in characters that people can read.	Lesk (1995)	N/A
2	No Definite Term	XML represents a format that is easy to restore and understand by both humans and machines.	Müller et al. (2003)	N/A
2	No Definite Term	XML is an open and established notation.	Müller et al. (2003)	N/A
2	No Definite Term	To allow data to be shared across information systems and remain impervious to many proprietary software revisions.	Potter (2006)	OpenOffice (Yes)

2	No Definite Term	Inability to be backed out into a usable format.	Potter (2006)	PDFs (No)
2	Dependencies	The characteristics (not dependent on specific hardware, not dependent on specific operating systems, not dependent on one specific reader and not dependent on other external resources) indicate the dependency on a specific environment or other resources such as fonts and codecs.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Limited) MS Word (Little)
2	Adoption	The characteristics (worldwide usage and usage in the cultural heritage sector as archival format) indicate the popularity and ubiquity of a file format.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Yes) MS Word (Limited)
2	Component Reuse	Not Defined	Sahu (2006)	PDF (No) HTML (Limited) SGML (Excellent)
2	Data Interchange	Not Defined	Sahu (2006)	PDF (No) HTML (Limited) SGML (Excellent)
2	Re-purposing	Not Defined	Sahu (2006)	PDF (Limited) HTML (Limited) SGML (Excellent)
2	Open Standard	Not Defined	Sahu (2006)	N/A
2	Ubiquity	Not Defined	Sahu (2006)	N/A
2	Stability	Not Defined	Sahu (2006)	N/A
2	Viability	Not Defined	Sahu (2006)	N/A
2	Transparency	Level A conforming PDF/A files provide text “in natural reading order” so that the file can be read with basic text editing tools such as MS Notepad.	Suillivan (2006)	PDF/A (Yes) MS Notepad (Yes)

2	Adoption	Designed for flexibility of implementation to promote its wide adoption.	Suillivan (2006)	PDF/A (Yes)
3	Transparency	Amenable to direct analysis with basic tools.	Abrams et al. (2005)	N/A
3	Authenticity	The format must preserve the content (data and structure) of the record, and any inherent contextual, provenance, referencing and fixity information.	Brown (2003)	N/A
3	Presentation	If the authenticity of an electronic record requires preservation of its original „look and feel" (fonts, colors and layout), then the ability of a file format to support this through migration will be a crucial consideration.	Brown (2003)	N/A
3	Integrity of Layout	Not Defined	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (N/A)
3	Integrity of Rendering of Equations	Not Defined	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (N/A)
3	Integrity of Structure	Not Defined	CENDI (2007) Hodge & Anderson (2007)	PDF (Limited) PDF/A (Limited) TIFF_G4 (N/A)
3	Provenance Traceability	Ability to trace the entire configuration of data production – based on information in the files and n the documentation of how the data were produced.	Folk & Barkstrom (2003)	N/A
3	Citability	A machine-independent ability to reference or "cite" the individual data element in a stable way.	Folk & Barkstrom (2003)	N/A
3	Referential Extensibility	The ability to build annotations about new interpretations of the data – and to preserve those annotations. This gives us the ability to create indexes of interesting phenomena that are external to the original files.	Folk & Barkstrom (2003)	N/A

3	Source Verification	Cryptographic encoding of files or digital watermarks must be created without overburdening the data centers or archives.	Folk & Barkstrom (2003)	N/A
4	Technical Protection Mechanisms	No encryption, passwords, etc.	Abrams et al. (2005)	N/A
4	No Definite Term	The eXtensible Characterisation Languages (XCL) support the automatic validation of document conversions and the evaluation of conversion quality by hierarchically decomposing documents from different sources and representing them in an abstract XML language.	Becker et al. (2008a) Becker et al. (2008b)	N/A
4	Feature Set	Formats supporting the full range of features and functionality required for their designated purpose or business process.	Brown (2003)	N/A
4	Viability	Some formats provide error-detection facilities, to allow detection of file corruption which may have occurred during transmission. Many formats include a CRC (Cyclic Redundancy Check) value for this purpose.	Brown (2003)	PNG format (Yes)
4	Technical Protection Mechanism	Implementation of a mechanism that prevents the preservation of content by a trusted authority.	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (No) TIFF_G4 (No)
4	Normal Rendering	Not Defined	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Limited) TIFF_G4 (Yes)
4	Beyond Normal Rendering	Not Defined	CENDI (2007) Hodge & Anderson (2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes)
4	Support for Graphic Effects and Typography	Not Defined	CENDI (2007) Hodge & Anderson (2007)	TIFF_G4 (No)
4	Color Maintenance	Not Defined	CENDI (2007) Hodge & Anderson (2007)	TIFF_G4 (Limited)

4	Clarity	Support for high image resolution.	CENDI (2007) Hodge & Anderson (2007)	TIFF_G4 (Yes)
4	Markup Compatibility and Extensibility	To support a much broader range of applications.	ECMA (2008)	N/A
4	Compactness	To minimize storage and I/O costs.	Folk & Barkstrom (2003)	N/A
4	Size	Access to digital objects, especially objects stored on tape; necessarily incurs some overhead due to latency. One way to overcome this latency is to transfer data in large blocks.	Folk & Barkstrom (2003)	N/A
4	Ability to Aggregate Many Objects in a Single File.	A file format that supports the aggregation of many digital objects in one file can enable an archive to maintain as small of an archive “name space” as possible.	Folk & Barkstrom (2003)	N/A
4	URN Embedding Capability	Files could reference documentation or link with other files.	Folk & Barkstrom (2003)	N/A
4	File Corruption Detection	Being able to detect that a file has been corrupted. Corruption detection is useful not only for protecting against malicious actions, but also against unintended changes in the data, such as that caused by faulty equipment.	Folk & Barkstrom (2003)	N/A
4	File Corruption Correction	To find ways of using error-detection approaches to provide error-correction.	Folk & Barkstrom (2003)	N/A
4	No Definite Term	XML is a human readable text format and internationalized character sets are supported.	Müller et al. (2003)	N/A
4	Complexity	The characteristics (human readability, compression, variety of features) indicate how complicated a file format can be to decipher.	Rog & Wijk (2007) Wijk & Rog (2007)	N/A

4	Robustness	The characteristics (robust against single point of failure, support for file corruption detection, file format stability, backward compatibility and forward compatibility) indicate the extent to which the format changes over time and the extent to which successive generations differ from each other and provide information on the ways the file format is protected against file corruption.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Limited) MS Word (Limited)
4	Technical Protection Mechanism (DRM)	The characteristics (password protection, copy protection, digital signature, printing protection and content extraction protection) indicate the possibilities in a file format to restrict access (in a broad sense) to content.	Rog & Wijk (2007) Wijk & Rog (2007)	PDF/A-1 (Limited) MS Word (Limited)
4	Feature Set	Not Defined	Sahu (2006)	N/A
4	Distributing Page Image	Not Defined	Sahu (2006)	PDF (Excellent) HTML (Good) SGML (Good)
4	Searching	Not Defined	Sahu (2006)	PDF (Limited) HTML (Good) SGML (Excellent)

5. Criteria of File Formats in Four Families

5.1 Autonomy Family

Autonomy refers to independence from outside proprietary or commercial control (Stanescu, 2005). Autonomy of the file format refers to several factors. First, the document should be self-contained, meaning the content information (e.g., the text), the structural information (i.e., for those documents that are structured), the formatting information (e.g., fonts, colors, styles, etc.), as well as the metadata information. The lack of a self-contained format may also be a problem for archivists (Sullivan, 2006; Hodge & Anderson, 2007). Self-containment does not necessarily mean that archivists will only have one document to deal with. It does, however, mean that they will have documents that will provide them with all the information to access and process the content, the structure, the formatting and the metadata. Another factor is the independence of the document from proprietary or commercial hardware and software configurations, especially to prevent any issues with software versions, outdated material or patent and copyright issues. Ideally, a simple text editor, reader or browser should support this format, such as Adobe Reader, which is freely available on the Internet, or the text editor supplied with the operating system of a computer such as Notepad on Windows-based computers and TextEdit on Apple-based computers. Having documents in a proprietary format controlled by a third party means that at some point in the future this format may no longer be supported, or that a change in the user agreement may lead to restricted access. This independence also means that the document must be freely accessible, without password restriction or protection, and without any digital rights management scheme (these criteria are prescriptions with PDF/A). Restricting access to a document with a password can lead to serious problems if the password gets lost. By definition, access restriction is the antithesis of long-term preservation. Finally, the size and compactness of the document will influence the selection of a file format. Examples of criteria to be considered in the autonomy family are: metadata support, self-documentation, openness, open availability, dependencies, device independencies, external dependency, etc.

5.2 Interoperability Family

Interoperability refers primarily to the ability of a file format to be compatible with other formats and exchange documents without loss of information (The National Archives, 2003; ECMA, 2006). More specifically, it refers to the ability of a given type of software to open a

document without requiring any special application, plug-in, codec, or proprietary add-on. Adherence to open source standards is usually a good indication of the interoperability of a format. Usually, an open standard is released after years of bargaining and agreements between major players. Supervision by an international standard (such as ISO or the W3C) usually helps propagate the format. Some good examples of open standard formats include HTML, XML, SMIL, SVG (the latter two of which are subsets of XML and handle multimedia integration and presentations and vector graphics respectively). Examples of criteria to be considered for the interoperability family are robustness, data interchange, etc.

All of these formats are official recommendations by the W3C. TIFF and PDF/A are also both open standards and published by ISO. Of course, for standards to be interoperable, they need to achieve a certain level of maturity and ubiquity. The more a standard is used, the more it is likely others will develop other standards compatible to it. TIFF and PDF are widely used, even by the general public. XML is another good example of this, with standards as varied as SMIL, XHTML (the XML version of HTML), LaTeX (typesetting of mathematical formulas), RDF (used among other things in bibliographic records), EAD (an archival description format), etc. A priori, all these XML-derived standards are compatible, either natively or through a conversion/mapping process, assuming that what these standards describe can be compared. Practical applications of XML standards are in exchange information protocols such as RSS, or multi-purposing to enable the ability to "package" the same content for consumption on a personal computer or laptop, on a wireless device, or a cell phone.

5.3 Authenticity Family

Authenticity refers to the ability to guarantee that a file is what it originally was, without any corruption or alteration, and represents the content accurately (Becker et al, 2008; The National Archives, 2003). Specific to authenticity is data integrity, which assesses the integrity of the file through an internal mechanism (e.g., PNG files include byte sequences to validate against errors). Another method to validate the authenticity of a document is to look at its traceability (Folk & Barkstrom, 2003); that is, the traces left by the original author, those who modified a file, those who opened a file, etc. One basic example is the difference between the creation date, modification date and access date of any file on a personal computer. These three dates correspond to a moment when someone, often not the same person, opened and/or used the file. Other mechanisms may require log information, which are external to the file. Another good

indication of authenticity is the stability of a format (The National Archives, 2003; Rog & van Wijk, 2007). Examples of criteria to be considered for the authenticity family are: integrity of layout, integrity of structure, etc. A format that is widely used is more likely to be stable. A stable format is also more likely to cause less data loss and data corruption, thus being a better indicator of authenticity.

5.4 Functionality Family

Functionality refers to the ability of a format to do exactly what it is supposed to be technically doing. This is why it is important to distinguish between two broad uses: preservation of the document structure and formatting, and preservation of useable content. Examples of criteria to be considered in the functionality family are: technical protection mechanism, adoption, component reuse, etc. To preserve the formatting of a document, one needs to create a static copy, a "published view" of a given content. This may be suitable for documents intended for distribution. Other content, such as database information or device specific documents, may not necessarily be well suited for such a preservation method. In these cases, preserving the content is what is more important. It may be more important to preserve the layout of a marketing brochure than its content, whereas it may be more important to preserve the content of financial statements than their formatting. The decision to preserve one over the other will rest with the author, the records manager or the archivist, and a file format will need to be chosen to better suit that need.

6. Discussion and Conclusion

The objectives of this research were to examine the various criteria described in the literature in the major families and identify the basic requirements of the appropriate file formats for long-term preservation, and to access and help select the best file formats. We have identified the major criteria that are used in assessing file formats.

Going back to the beginning of this study, the open standard format is defined as "formats for which the technical specifications have been made available in the public domain" (The National Archives, 2003). It refers to independence from outside proprietary or commercial control (Stanescu, 2005). Open standard image file formats "are widely accepted, have freely available specifications, are highly interoperable, incorporate no data compression and are

capable of supporting preservation metadata" (Horik, 2004). Coyle (2002) explains the three characteristics of open standards: 1) anyone can use the standards to develop software, 2) anyone can acquire the standards for free or without a significant cost and 3) the standard has been developed in such a way that anyone can participate. The open standard is related to *open access*, which comes from the Open Access movement that allows resources to be freely available to the public and permits any user to use those resources (e.g., electronic journals, repositories, databases, software applications, etc.) without financial, legal, or technical barriers to their access. Since the 1990s, as the term has been broadly adopted in many fields and has become prevalent in entire library communities, open access to resources is a useful way to provide users with better services. According to these characteristics, the XML file format seems mostly open, although it has many subsets with different technical specifications that make them dependent on a specific file provider. Which file formats are most appropriate to archivists, records managers and users? How can we examine which file formats are proper for long-term preservation and persistent access? These questions remain the choice of a user. We hope that a close examination of the criteria of file formats can help archivists and librarians choose appropriate file formats for the long-term preservation of electronic documents.

7. References

Adobe (2010). "Adobe fast facts." Retrieved November 1, 2011 from http://www.adobe.com/aboutadobe/pressroom/pdfs/fastfacts.pdf.

Barnes, I. (2006). "Preservation of word processing documents." Retrieved November 1, 2011 from http://www.apsr.edu.au/publications/word_processing_preservation.pdf.

Becker C., Rauber A., Heydegger V., Schnasse J., and Thaller M. (2008). "A generic XML language for characterising objects to support digital preservation." In *Proceedings of the 2008 ACM symposium on Applied computing, March 16-20, 2008, Fortaleza, Ceara, Brazil*. Retrieved November 1, 2011 from http://www.apsr.edu.au/publications/word_processing_preservation.pdf.

CENDI Digital Preservation Task Group (2007), "Formats for digital preservation: A review of alternatives and issues." Retrieved November 1, 2011 from http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf.

Coyle, K. (2002). "Open Source, Open Standards," *Information Technology and Libraries*, 21, 1, 33-36.

DAVID 4 Project (Digital ArchiVing, guIdeline and aDvice 4) (2003). "Standards for

Fileformats," 1. Retrieved November 1, 2011 from http://www.expertisecentrumdavid.be/davidproject/teksten/guideline4.pdf.

ECMA International (2006). "Office Open XML file formats – ECMA-376 1st edition Part 1." Retrieved November 1, 2011 from http://www.ecma-international.org/publications/standards/Ecma-376.htm.

Folk, M., and Barkstrom, B.R. (2003). "Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries." Paper presented at the Joint Conference on Digital Libraries (JCDL), Houston, TX. Retrieved November 1, 2011 from http://ftp.hdfgroup.org/projects/nara/Sci_Formats_and_Archiving.pdf.

Hodge, G., and Anderson, N. (2007). "Formats for digital preservation: A review of alternatives and issues," *Information Services & Use*, 27, 45-63.

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).

ISO/IEC 26300:2006. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485.

Lesk, M. (1995). "Preserving digital objects: Recurrent needs and challenges." In *Proceedings of the 2nd NPO Conference on Multimedia Preservation, Brisbane*. Retrieved November 1, 2011 from http://www.lesk.com/mlesk/auspres/aus.html.

Microsoft Corporation (2010). "About indexing text in TIFF and MDI files." Retrieved November 1, 2011 from http://office.microsoft.com/en-us/help/HP030812361033.aspx?pid=CH010000951033.

Office Open XML (2005). http://www.microsoft.com/presspass/features/2005/nov05/11-21Ecma.mspx and ISO/IEC DIS 29500:2008 and http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51463.

OpenDocument. (2009, Novermber 27). In Wikipedia, the free encyclopedia. Retrieved November 1, 2011 from http://en.wikipedia.org/wiki/OpenDocument.

PDF. (2009, December 4). In Wikipedia, the free encyclopedia. Retrieved November 1, 2011 from http://en.wikipedia.org/wiki/PDF.

Potter, J.M., (2006). "Formats Conversion technologies set to benefit institutional repositories." Retrieved November 1, 2011 from http://www.freewebs.com/academicportfolio/formatpaper.pdf.

Rog, J., and Wijk, V.C. (2007). "Evaluating File Formats for Long-term Preservation." Retrieved November 1, 2011 from http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_2702208.pdf.

Stanescu, A. (2005). "Assessing the durability of formats in a digital preservation environment. The INFORM methodology," *OCLC Systems & Services*, 21(1), 61-81.

Suillivan, S.J. (2006). "An archival/records management perspective on PDF/A," *Records Management Journal, 16*(1), 51-56.

The National Archives. (2003). "Selecting file formats for long-term preservation." Retrieved November 1, 2011 from http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf.

van Horik, R. (2004). "Image Formats: Practical Experiences" (paper presented in Erpanet Training, Vienna, May 10-11, 2004), 22. Retrieved November 1, 2011 from http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Horik.pdf.

Wijk, V.C. and Rog, J. (2007). "Evaluating File Formats for Long-term Preservation; PPT presentation." Retrieved November 1, 2011 from http://ipres.las.ac.cn/pdf/Caroline-iPRES2007-11-12oct_CW.pdf.

World Wide Web Consortium. (2008). "Extensible markup language (XML) 1.0 (fifth edition) – W3C recommendation 26 November 2008." Retrieved November 1, 2011 from http://www.w3.org/TR/REC-xml.

World Wide Web Consortium. (2009). "Facts about W3C." Retrieved November 1, 2011 from http://www.w3.org/Consortium/facts.

8. Studies Included in the Table

Abrams, S., Fanning, B., Helander, D., and Sullivan, S. (2005). *PDF-A: The Development of a Digital Preservation Standard.* Retrieved November 1, 2011 from http://www.aiim.org/documents/standards/PDF-A.ppt.

Barnes, I. (2006). "Preservation of word processing documents." Retrieved November 1, 2011 from http://www.apsr.edu.au/publications/word_processing_preservation.pdf.

Becker C., Rauber A., Heydegger V., Schnasse J., and Thaller M. (2008a). "A generic XML language for characterising objects to support digital preservation." In *Proceedings of the 2008 ACM Symposium on Applied Computing, March 16-20, 2008, Fortaleza, Ceara, Brazil*.

Becker C., Rauber A., Heydegger V., Schnasse J., and Thaller M. (2008b). "Systematic Characterization of Objects in Digital Preservation: The eXtensible Characterization Language," *Journal of Universal Computer Science*, 14 (18), 2936-2952.

Brown, A. (2003). "The National Archives. Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation." Retrieved November 1, 2011 from http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf.

CENDI Digital Preservation Task Group (2007), "Formats for digital preservation: A review of alternatives and issues." Retrieved November 1, 2011 from http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf.

ECMA (2008), "Standard ECMA-376: Office Open XML File - part 1." Retrieved November 1, 2011 from http://www.ecma-international.org/publications/standards/Ecma-376.htm.

File format. (2009, December 16). In Wikipedia, the free encyclopedia. Retrieved November 1, 2011 from http://en.wikipedia.org/wiki/File_format.

Folk, M., and Barkstrom, B.R. (2003). "Attributes of File Formats for Long-Term Preservation of Scientific and Engineering Data in Digital Libraries." Paper presented at the Joint Conference on Digital Libraries (JCDL), Houston, TX.

Hodge, G., and Anderson, N. (2007). "Formats for digital preservation: A review of alternatives and issues," *Information Services & Use*, 27, 45-63.

Johnson, A.H. (Oct 18, 1999). "XML Xtends its reach: XML finds favor in many IT shops, but it's still not right for everyone," *Computerworld, 33*(42), 76-81.

Lesk, M. (1995). "Preserving digital objects: Recurrent needs and challenges." In *Proceedings of the 2nd NPO Conference on Multimedia Preservation, Brisbane*. Retrieved November 1, 2011 from http://www.lesk.com/mlesk/auspres/aus.html.

Müller, E., Klosa, U., Hansson, P., Andersson, S. and Siira, E. (2003). "Using XML for Long- Term Preservation: Experiences from the DiVA Project." In *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations*, May 2003, 109-116.

OpenDocument. (2009, Novermber 27). In Wikipedia, the free encyclopedia. Retrieved November 1, 2011 from http://en.wikipedia.org/wiki/OpenDocument.

PDF. (2009, December 4). In Wikipedia, the free encyclopedia. Retrieved November 1, 2011 from http://en.wikipedia.org/wiki/PDF.

Potter, J.M., (2006). "Formats Conversion technologies set to benefit institutional repositories." Retrieved November 1, 2011 from http://www.freewebs.com/academicportfolio/formatpaper.pdf.

Rog, J., & Wijk, V.C. (2007). "Evaluating File Formats for Long-term Preservation." Retrieved November 1, 2011 from https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf.

Sahu, D.K. (2006). "Long term preservation: which file format to use." In *Workshops on Open Access & Institutional Repository, 2-8 May 2004, Chennai, India*. Retrieved November 1, 2011 from http://openmed.nic.in/1363/01/Long_term_preservation.pdf.

Suillivan, S.J. (2006). "An archival/records management perspective on PDF/A," *Records Management Journal, 16* (1), 51-56.

Wijk, V.C. and Rog, J. (2007). "Evaluating File Formats for Long-term Preservation." Retrieved November 1, 2011 from http://ipres.las.ac.cn/pdf/*Caroline-iPRES2007-11-12oct_CW*.pdf.

9. Research Team

The InterPARES Project would like to thank the researchers and research assistants that contributed to the development of General Study 20:

Researchers:
Eun G. Park, TEAM Korea
Sam Oh, Director of TEAM Korea

Graduate Research Assistant:
Dongwook Kim