

Examining Attributes of Open Standard File Formats for Long-term Preservation and Open Access

Eun G.Park and Sam Oh

ABSTRACT

This study examines the attributes that have been used to assess file formats in literature and compiles the most frequently used attributes of file formats to establish open-standard file-format-selection criteria. A comprehensive review was undertaken to identify the current knowledge regarding file-format-selection criteria. The findings indicate that the most common criteria can be categorized into five major groups: functionality, metadata, openness, interoperability, and independence. These attributes appear to be closely related. Additional attributes include presentation, authenticity, adoption, protection, preservation, reference, and others.

INTRODUCTION

File format is one of the core issues in the fields of digital content management and digital preservation. As many different types of file formats are available for texts, images, graphs, audio recordings, videos, databases, and web applications, the selection of appropriate file formats poses an ongoing challenge to libraries, archives, and other cultural heritage institutions. Some file formats appear to be more widely accepted: Tagged Image File Format (TIFF), Portable Document Format (PDF), PDF/A, Office Open XML (OOXML), and Open Document Format (ODF), to name a few. Many institutions, including the Library of Congress (LC), possess guidelines on file format applications for long-term preservation strategies that specify requisite characteristics of acceptable file formats (e.g., they are independent of specific operating systems, are independent of hardware and software functions, conform to international standards, etc.).¹ The Format Descriptions database of the Global Digital Format Registry is an effort to maintain a detailed representation of information and sustainability factors for as many file formats as possible (the PRONOM technical registry is another such database).² Despite these developments, file format selection remains a complex task and prompts many questions that range from a general interest (“Which selection criteria are appropriate?”) to more specific (“Are these international standard file formats sufficient for us to ensure long term preservation and access?” or “How should we define and implement standard file formats in harmony with our local context?”).

In this study, we investigate the definitions and features of standard file formats and examine the

Eun G. Park (eun.park@mcgill.ca) is Associate Professor, School of Information Studies, McGill University, Montreal, Canada. **Sam Oh** (samoh@skku.edu) is corresponding author and Professor, Department of Library and Information Science, Sungkyunkwan University, Seoul, Korea.

major attributes of assessing file formats. We discuss relevant issues from the viewpoint of open-standard file formats for long-term preservation and open access.

BACKGROUND ON STANDARD FILE FORMATS

The term *file format* is generally defined as what “specifies the organization of information at some level of abstraction, contained in one or more byte streams that can be exchanged between systems.”³ According to InterPARES 2, file format is “the organization of data within files, usually designed to facilitate the storage, retrieval, processing, presentation, and/or transmission of the data by software.”⁴ The PREMIS Data Dictionary for Preservation Metadata observes that, technically, file format is “a specific, pre-established structure for the organization of a digital file or bitstream.”⁵

In general, file format can be divided into two types: an access format and a preservation format. An access format is “suitable for viewing a document or doing something with it so that users access the on-the-fly converted access formats.”⁶ In comparison, a preservation format is “suitable for storing a document in an electronic archive for a long period”⁷; it provides “the ability to capture the material into the archive and render and disseminate the information now and in the future.”⁸ While the ability to ensure long-term preservation focuses on the sustainability of preservation formats, the document in its access format tends to emphasize that it should be accessible and available by users, presumably all of the time.

Many researchers have discussed file formats and long-term preservation in relation to various types of resources. For example, Folk and Barkstrom describe and adopt several attributes of file formats that may affect the long-term preservation of scientific and engineering data (e.g., the ease of archival storage, ease of archival access, usability, data scholarship enablement, support for data integrity, and maintainability and durability of file formats).⁹ Barnes suggests converting word processing documents in digital repositories, which are unsuitable for long-term storage, into a preservation format.¹⁰ The evaluation by Rauch, Krottmaier, and Tochtermann illustrates the practical use of file formats for 3D objects in terms of long-term reliability.¹¹

Others have developed and/or applied numerous criteria in different settings. For instance, Sullivan uses a list of desirable properties of a long-term preservation format to explain the purpose of PDF/A from an archival and records management perspective.¹² Sullivan cites device independence, self-containment, self-describing, transparency, accessibility, disclosure, and adoption as such properties. Rauch, Krottmaier, and Tochtermann’s study applies criteria that consist of technical characteristics (e.g., open specification, compatibility, and standardization) and market characteristics (e.g., guarantee duration, support duration, market penetration, and the number of independent producers). Rog and van Wijk propose a quantifiable assessment method to calculate composite scores of file formats.¹³ They identify seven main categories of criteria: openness, adoption, complexity, technical protection mechanism, self-documentation, robustness, and dependencies. Sahu focuses on the criteria developed by the UK’s National Archives, which include open standards, ubiquity, stability, metadata support, feature set,

interoperability, and viability.¹⁴ A more comprehensive evaluation by the LC reveals three components—technical factors, quality, and functionality—while placing a particular emphasis on the balance between the first two.¹⁵ Hodge and Anderson use seven criteria for sustainability, which are similar to the technical factors of the LC study: disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms.¹⁶

Some institutions adopt another term, *standard file formats*, to differentiate accepted and recommended file formats from others. According to the DAVID project, “standard file formats owe their status to (official) initiatives for standardizing or to their widespread use.”¹⁷ *Standard* may be too general to specify the elements of file formats. However, there is a recognition that only those file formats accepted and recommended by national or international standard organizations (such as the International Standardization Organization [ISO], International Industry Imaging Association [I3A], WWW Consortium, etc.) are genuine standard file formats. For example, ISO has announced several standard file formats for images: TIFF/IT (ISO 12639:2004), PNG (ISO/IEC 15948:2004), and JPEG 2000 (ISO/IEC 15444:2003, 2004, 2005, 2007, 2008). For document file formats, PDF/A-1 (ISO Standard 19005-1. *Document File Format for Long-Term Preservation*) is one example. This format is proprietary to maintain archival and records-management requirements and to preserve the visual appearance and migration needs of electronic documents. Office Open XML file format (ISO/IEC 29500-1:2008. Information Technology—Document Description and Processing Languages) is another open standard that can be implemented from Microsoft Office applications on multiple platforms. ODF (ISO/IEC 26300:2006. Information Technology—Open Document Format for Office Applications [OpenDocument] v1.0) is an XML-based open file format. Regardless of ISO-announced standards, some errors in these file formats have been reported. For example, although PDF/A-1 is for long-term preservation of and access to documents, studies reveal that the feature-rich nature of PDF can create difficulties in preserving PDF information over time.¹⁸ To overcome the barriers of PDF and PDF/A-1, XML technology seems prevalent for digital resources in archiving systems and digital preservation.¹⁹ The digital repository community is treating XML technology as a panacea and converting most of their digital resources to XML.

The Netherlands Institute for Scientific Information Service (NISIS) adopts another noteworthy definition of standard file formats. It observes that 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.”²⁰ This definition implies specific and advanced ramifications for cost-free interoperability and metadata, which closely relates to open access.

Open standard is another relevant term to consider in file formats. Although perspectives vary greatly between researchers, open standards can be acquired and used without any barrier or cost.²¹ In other words, open standard products are free from restrictions, such as patents, and are independent of proprietary hardware or software. Since the 1990s, open standard has been broadly adopted in many fields and is now an almost compulsory feature in information services.

To follow the National Archives' definition, open standard formats are "formats for which the technical specifications have been made available in the public domain."²² In comparison, the Folk and Barkstrom approach opens standards from institutional support perspectives, relying on user communities for standards that are widely available and used.²³ On a more specific level, Stanescu emphasizes independence as the basic selection criteria for file formats.²⁴ Others, such as Todd, propose determining whether a standard should be more open than others by applying criteria: adoption, platform independence, disclosure, transparency, and metadata support.²⁵ Other factors considered by Todd include reusability and interoperability; robustness, complexity, and viability; stability; and intellectual property (IP) and rights management.²⁶ Echoing the LC, Hodge and Anderson also suggest a list of selection criteria that have been grouped under the banner of "technical factors": disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms.²⁷

Researchers agree that open standard file formats are less obsolete and more reliable than proprietary formats.²⁸ Close examination of the NISIS definition mentioned above reveals that standard file formats are in reality not free, nor do they allow unrestricted access to resources. The three file formats that ISO has announced (PDF/A, OOXML, and ODF) are proprietary and sometimes costly. They also prohibit the purchase of access to a proprietary standard, although there is an assumption that a standard should be free from legal and financial restrictions. The ISO-announced file formats, in short, are only standard file formats, not *open* standard file formats.

For cultural heritage institutions, questions regarding appropriate selection criteria and the sufficiency of existing international standard file formats for long-term preservation and access remain unanswered. There exists neither a uniform method to compare the specifications of different file formats nor an objective approach to assess format specifications that would ensure long-term preservation and persistent access.

OBJECTIVES OF THE STUDY

In this study, we attempt to better define and establish open-standard file-format-selection criteria. To that end, we assess and compile the most frequently used attributes of file formats to establish open-standard file-format-selection criteria.

METHOD

We performed a comprehensive review of published articles, institutional reports, and other literature to identify the current knowledge regarding file-format-selection criteria. We included literature that deals with the three standard file formats (PDF, PDF/A, and XML) but excluded the recently announced ODF format due to the scarcity of literature on ODF. Among more than the thirty articles initially reviewed, only twenty-five that use their own clear attributes were included in this study. All of the attributes that we have employed are listed by frequency and grouped according to similarities in meaning (see appendix). The original definitions or descriptions that we used are listed in the second column. The file formats that we assessed by their attributes are

listed in the third column. When we give attributes without specific definitions or descriptions, “no definite term” is inserted.

FINDINGS

As illustrated in the appendix, the criteria identified by the studies vary. Although the requirements and context of the studies may differ, the most common criteria can be divided into five categories: functionality, metadata, openness, interoperability, and independence.

First, functionality refers to the ability of a format to do exactly what it is supposed to be doing.²⁹ It is important to distinguish between two broad uses: preservation of document structure and formatting and preservation of useable content. To preserve document formatting, a “published view” of a given piece of content is critical for distribution. Other content, such as database information or device-specific documents, needs to be preserved as well. Functionality criteria include various attributes related to formats and structure or physical and technical specifications of files (e.g., robustness, feature set, viability, color maintenance, clarity, compactness, modularity, compression algorithms, etc.).

Second, metadata indicates that a format allows rich descriptive and technical metadata to be embedded in files. Metadata can be expressed as metadata support, self-documentation (self-documenting), documentation, content-level (as opposed to presentation-level) description, self-describing, self-describing files, formal description of format, etc.

Third, openness refers to specifications of a file format that are publicly available and accessible and formats that are not proprietary. Whether seen as a single definition or as a set of criteria, the characteristic that appears to be at the core of the open standard movement is its independence from outside proprietary or commercial control. Openness also may refer to the autonomy of a file format, which relies on several factors. First, the document should be self-contained in terms of the content information (e.g., the text), the structural information (i.e., for those documents that are structured), the formatting information (e.g., fonts, colours, styles, etc.), and the metadata information. Self-containment does not necessarily mean that an archivist will only have one document to deal with. It does mean, however, that they will have documents that will provide them with all the information to access and process the content, structure, formatting, and metadata. Openness is expressed as *open availability* by some researchers.³⁰ Other researchers adopt the term *disclosure* for expressing that specification is publicly available.³¹

Fourth, is the independence of a document from proprietary or commercial hardware and software configurations, especially to prevent any issues resulting from different versions of software, hardware, and operating systems. This aspect is expressed in the appendix as open standards, open source software or equivalent, standard/proprietary, etc. This also closely relates to independence, one of the five categories in the appendix, expressed as device independencies, independent implementations, no external dependency, no external dependencies, portability, and monitoring obsolescence. Having documents in a proprietary format controlled by a third party

implies that, at one time or another, this format may no longer be supported, or that a change in the user agreement may lead to restricted access, access to outdated material, or patent and copyright issues. This fact means that the document must be freely accessible, without password restrictions or protection, and without any digital rights management scheme. Blocking access to a document with a password can lead to serious problems if the password gets lost. In addition, the size and compactness of the document will influence the selection of a file format.

Fifth, interoperability primarily refers to the ability of a file format to be compatible with other formats and to exchange documents without loss of information.³² Specifically, it refers to the ability of a given 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. In general, 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) commonly helps propagate the format.

In addition to the five categories mentioned above, other attributes are often used. Presentation, authenticity, adoption, protection, preservation and reference are such examples. Among these attributes, authenticity, although this is the seventh in the appendix, is one of the most important attributes in archives and records management. It refers to the ability to guarantee that a file is what it originally was without any corruption or alteration.³³ 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 of validating the authenticity of a document is to look at its traceability,³⁴ that is, the traces left by the original author and those who modified or opened a file. One 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 a different person each time) opened the file. Other mechanisms may require log information, which is external to the file. Another good indication of authenticity is the stability of a format.³⁵ 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 corruption; hence it is a better indicator of authenticity. Presentation includes attributes related to presenting and rendering data, expressed as *distributing a page image*, *normal rendering*, *self-containment*, *self-contained*, and *beyond normal rendering*. Adoption indicates how popular and widely a file format is adopted by user communities, also represented as *popularity*, *widely used formats*, *ubiquity*, or *continuity*. Protection includes the technical protection mechanism or source verification to protect with security skills. Preservation means long-term preservation, institutional support, or ease of transformation and preservation. Reference indicates citability, or referential extensibility. Among other attributes, transparency is interesting to note because it indicates the degree to which files are open to direct analysis with basic tools and human readability.

Another important aspect across these criteria is that the terminologies used in the studies may be quite different yet describe the same or similar concepts from different angles. For instance, Rog and van Wijk use *openness* for standardization and specification without restrictions,³⁶ while

several other researchers use *open availability* to convey the same thing.³⁷ They in turn adopt the term *disclosure* to express that specification is publicly available.³⁸

DISCUSSION AND CONCLUSION

Functionality, metadata, openness, interoperability, and independence appear to be the most important factors when selecting file formats. When file formats for long-term preservation and open access are under discussion, cultural heritage institutions need to consider many issues. Despite several efforts, it is still tricky for them to identify the most appropriate file format or even to discern acceptable formats from unacceptable formats. Where it is difficult to prevent the creation of a new file format, format selection is not an easy task, both in theory and in practice. It is critical, however, to base the decision on a clear understanding of the purpose for which the document is preserved: access preservation or repurposing preservation. Cultural heritage institutions and digital repository communities need to guarantee long-term preservation of digital resources in selected file formats. Additionally, users find it necessary to have access to digital information in these file formats. Additional consideration involves the level of access users may enjoy (e.g., long-term access, permanent access, open access, persistent access, etc.).

When determining international standard file formats, an aspect of open access should be included because it is a well-liked topic. It is necessary to develop a scale or measurement to assess open-standard format specifications to ensure long-term preservation and open access. Identifying which attributes are required to be an open-standard file format and which digital format is most apt for the use and sustainability of long-term preservation is a meaningful task.

The outcome of our study provides a framework for appropriate strategies when selecting file formats for long-term preservation and access to digital content. We hope that the criteria described in this study will benefit librarians, preservers, record creators, record managers, archivists, and users. We are reminded of Todd's remark that "the most important action is to align the recognition and weighting of criteria with a clear preservation strategy and keep them under review using risk management techniques."³⁹ The question of how to adopt and implement these attributes can only be answered in the local context and decisions of each cultural heritage institution.⁴⁰ Each institution should consider implementing a file format throughout the entire life cycle of digital resources, with a holistic approach to managerial, technical, procedural, archival, and financial issues for the purpose of long-term preservation and persistent access.

The criteria may change over time, as is necessary for any format to adequately serve its purpose. Maintaining its quality may be an ongoing task that cultural heritage institutions should take into account at all times. Even more importantly, cultural heritage institutions need to establish and implement a set of standard guidelines specific to each context for the selection of open-standard file formats.

Note: This research was supported by the Sungkyunkwan University Research Fund (2010-2011).

REFERENCES AND NOTES

1. Library of Congress, "Sustainability of Digital Formats: Planning for Library of Congress Collections," www.digitalpreservation.gov/formats/intro/intro.shtml (accessed November 21, 2011).
2. Global Digital Format Registry, www.gdfr.info (accessed November 17, 2011); The Technical Registry PRONOM, www.nationalarchives.gov.uk/aboutapps/pronom (accessed November 21, 2011).
3. Mike Folk and Bruce R. Barkstrom, "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, May 27–31, 2003), 1, www.larryblakeley.com/Articles/storage_archives_preservation/mike_folk_bruce_barkstrom200305.pdf (accessed November 21, 2011).
4. InterPARES 2 Project Glossary, p. 24, www.interpares.org/ip2/ip2_term_pdf.cfm?pdf=glossary (accessed November 21, 2011).
5. PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata*, ver. 2.0, March 2008, p. 195, www.loc.gov/standards/premis/v2/premis-2-0.pdf (accessed November 21, 2011).
6. Ian Barnes, "Preservation of Word Processing Documents," July 14, 2006, p. 4, http://apsr.anu.edu.au/publications/word_processing_preservation.pdf (accessed November 21, 2011).
7. Ibid.
8. Gail Hodge and Nikkia Anderson, "Formats for Digital Preservation: A Review of Alternatives and Issues," *Information Services & Use* 27 (2007): 46.
9. Folk and Barkstrom, "Attributes of File Formats."
10. Barnes, "Preservation of Word Processing Documents."
11. Carl Rauch, Harald Krottmaier, and Klaus Tochtermann, "File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats," in *Proceedings of the 11th International Conference on Electronic Publishing 2007* (Vienna, Austria, June 13–15, 2007): 101–6.
12. Susan J. Sullivan, "An Archival/Records Management Perspective on PDF/A," *Records Management Journal* 16, no. 1 (2006): 51–56.
13. Judith Rog and Caroline van Wijk, "Evaluating File Formats for Long-Term Preservation," 2008, www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (accessed November 21, 2011).

-
14. D. K. Sahu, "Long Term Preservation: Which File Format to Use" (paper presented in Workshops on Open Access & Institutional Repository, Chennai, India, May 2–8, 2004), http://openmed.nic.in/1363/01/Long_term_preservation.pdf (accessed November 21, 2011).
 15. CENDI Digital Preservation Task Group, "Formats for Digital Preservation: A Review of Alternatives and Issues," www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf (accessed November 21, 2011).
 16. Hodge and Anderson, "Formats for Digital Preservation."
 17. DAVID 4 Project (Digital Archiving, guideline and advice 4), "Standards for Fileformats," 1, www.expertisecentrumdavid.be/davidproject/teksten/guideline4.pdf (accessed November 21, 2011).
 18. Sullivan, "An Archival/Records Management Perspective on PDF/A"; John Michael Potter, "Formats Conversion Technologies Set to Benefit Institutional Repositories," <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.7881&rep=rep1&type=pdf> (accessed November 21, 2011).
 19. Eva Müller et al., "Using XML for Long-Term Preservation: Experiences from the DiVA Project," in *Proceedings of the 6th International Symposium on Electronic Theses and Dissertations* (May 20–24, 2003): 109–16, <https://edoc.hu-berlin.de/conferences/etd2003/hansson-peter/HTML/index.html> (accessed November 21, 2011).
 20. Rene van Horik, "Image Formats: Practical Experiences" (paper presented in Erpanet Training, Vienna, Austria, May 10–11, 2004), 22, www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Horik.pdf (accessed November 21, 2011).
 21. 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., mainly electronic journals, repositories, databases, software applications, etc.) without financial, legal, or technical barriers. See Amy E. C. Koehler, "Some Thoughts on the Meaning of Open Access for University Library Technical Services," *Serials Review* 32, no. 1 (March 2006): 17–21; Budapest Open Access Initiative, "Read the Budapest Open Access Initiative," www.soros.org/openaccess/read.shtml (accessed November 21, 2011).
 22. National Archives, "Selecting File Formats for Long-term Preservation," 6, www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_2_7022008.pdf (accessed November 21, 2011).
 23. Folk and Barkstrom, "Attributes of File Formats."

-
24. Andreas Stanescu, "Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology," *D-Lib Magazine* 10, no. 11 (November 2004), www.dlib.org/dlib/november04/stanescu/11stanescu.html (accessed November 21, 2011).
 25. Malcolm Todd, "Technology Watch Report: File Formats for Preservation," www.dpconline.org/advice/technology-watch-reports (accessed November 21, 2011).
 26. Ibid.
 27. Hodge and Anderson, "Formats for Digital Preservation."
 28. Edward M. Corrado, "The Importance of Open Access, Open Source, and Open Standards for Libraries," *Issues in Science & Technology Librarianship* (Spring 2005), www.library.ucsb.edu/istl/05-spring/article2.html (accessed November 21, 2011); Carl Vilbrandt et al., "Cultural Heritage Preservation Using Constructive Shape Modeling," *Computer Graphics Forum* 23, no. 1 (2004): 25–41; Marshall Breeding, "Preserving Digital Information," *Information Today* 19, no. 5 (2002): 48–49.
 29. Eun G. Park, "XML: Examining the Criteria to be Open Standard File Format," (paper presented at the InterPARES 3 International Symposium, Oslo, Norway, September 17, 2010), www.interpares.org/display_file.cfm?doc=IP3_ism04_presentation_3-3_korea.pdf (accessed November 21, 2011).
 30. Adrian Brown, "Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation," www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf (accessed November 21, 2011); Barnes, "Preservation of Word Processing Documents"; Sahu, "Long Term Preservation"; Potter, "Formats Conversion Technologies."
 31. Stephen Abrams et al., "PDF-A: The Development of a Digital Preservation Standard" (paper presented at the 69th Annual Meeting for the Society of American Archivists, New Orleans, Louisiana, August 14–21, 2005), www.aiim.org/documents/standards/PDF-A.ppt (accessed November 21, 2011); Sullivan, "An Archival/Records Management Perspective on PDF/A"; CENDI, "Formats for Digital Preservation"; and Hodge & Anderson, "Formats for Digital Preservation."
 32. The National Archives, http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (accessed November 21, 2011); ECMA International, "Office Open XML File Formats—ECMA-376," www.ecma-international.org/publications/standards/Ecma-376.htm (accessed November 21, 2011).
 33. Christoph Becker et al., "Systematic Characterisation of Objects in Digital Preservation: The Extensible Characterisation Languages," www.jucs.org/jucs_14_18/systematic_characterisation_of_objects/jucs_14_18_2936_2952_becker.pdf (accessed November 21, 2011); National Archives,

www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (accessed November 21, 2011).

34. Folk and Barkstrom, "Attributes of File Formats."
35. National Archives, www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (accessed November 21, 2011); Rog and van Wijk, "Evaluating File Formats for Long-term Preservation."
36. Rog and van Wijk, "Evaluating File Formats for Long-term Preservation."
37. See Brown, "Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation," www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf (accessed November 21, 2011); Barnes, "Preservation of Word Processing Documents"; Sahu, "Long Term Preservation"; Potter, "Formats Conversion Technologies."
38. Stephen Abrams et al., "PDF-A: The Development of a Digital Preservation Standard" (paper presented at the 69th Annual Meeting for the Society of American Archivists, New Orleans, Louisiana, August 14–21, 2005), www.aiim.org/documents/standards/PDF-A.ppt (accessed November 21, 2011).; Sullivan, "An Archival/Records Management Perspective on PDF/A"; CENDI, "Formats for Digital Preservation"; and Hodge & Anderson, "Formats for Digital Preservation."
39. Todd, "Technology Watch Report," 33.
40. Evelyn Peters McLellan, "Selecting Digital File Formats for Long-Term Preservation: InterPARES 2 Project General Study 11 Final Report," [www.interpares.org/display_file.cfm?doc=ip2_file_formats\(complete\).pdf](http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf) (accessed November 21, 2011).

APPENDIX: File Format Attributes

No.	Attribute	Definition/Description	Assessed File Format
1. F U N C T I O N A L I T Y	Robustness	Robust against single point of failure, support for file corruption detection, file format stability, backward compatibility and forward compatibility (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Limited) Microsoft Word (Limited)
		A robust format contains several layers of defense against corruption (Frey, 2000).	N/A
	Feature Set	Formats supporting the full range of features and functionality (Brown, 2003)	N/A
		Not defined (Sahu, 2006)	N/A
	Viability	Error-detection facilities to allow detection of file corruption (Brown, 2003).	PNG format (Yes)
		Not defined (Sahu, 2006)	N/A
	Support for Graphic Effects and Typography	Not defined (CENDI, 2007; Hodge & Anderson, 2007)	TIFF_G4 (No)
	Color Maintenance	Not defined (CENDI, 2007; Hodge & Anderson, 2007)	TIFF_G4 (Limited)
	Clarity	Support for high image resolution (CENDI, 2007; Hodge & Anderson, 2007)	TIFF_G4 (Yes)
	Quality	This pertains to how well the format fulfills its task today: (1) Low space costs, (2) highly encompassing, (3) robust, (4) simplicity, (5) highly tested, (6) loss-free, (7) supports metadata (Clausen, 2004).	N/A
	Compactness	To minimize storage and I/O costs (Folk & Barkstrom, 2003)	N/A
	Simplicity	Ease of implementing readers (Folk & Barkstrom, 2003)	N/A
	File Corruption Detection	To be able to detect that a file has been corrupted; to provide error-correction (Folk & Barkstrom, 2003)	N/A
	Raw I/O Efficiency	Formats that are organized for fast sequential access (Folk & Barkstrom, 2003)	N/A
	Availability of Readers	To maintain ease of data access for readers (Folk & Barkstrom, 2003)	N/A
	Ease of Subsetting	To process only part of data files (Folk & Barkstrom, 2003)	N/A
	Size	To transfer data in large blocks (Folk & Barkstrom, 2003)	N/A
	Ability to Aggregate Many Objects in a Single File	To maintain as small as archive "name space" as possible (Folk & Barkstrom, 2003)	N/A
	Ability to Embed Data Extraction Software in the Files	The files come with read software embedded (Folk & Barkstrom, 2003).	N/A
	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
Rigorous Definition	To be defined in a sufficient rigorous way (Folk & Barkstrom, 2003)	N/A	
Multilanguage Implementation of Library Software	To have multiple implementations of readers for a single format (Folk & Barkstrom, 2003)	N/A	
Memory	Some formats emphasize the presence or absence of memory (Frey, 2000).	TIFF (Yes)	

Accuracy	In some cases, the accuracy of the data can be decreased to save memory, e.g., through compression. In the case of a digital master, however, accuracy is very important (Frey, 2000).	N/A
Speed	The ability to access or display a data set at a certain speed is critical to certain applications (Frey, 2000).	N/A
Extendibility	A data format can be modified to allow for new types of data and features in the future (Frey, 2000).	N/A
Modularity	A modular data set definition is designed to allow some of its functionality to be upgraded or enhanced without having to propagate changes through all parts of the data set (Frey, 2000).	N/A
Plugability	Related to modularity, this permits the user of an implementation of a data set reader or writer to replace a module with private code (Frey, 2000).	N/A
Interpretability	Not binary formats (Barnes, 2006)	RTF (Yes) MS Word (No) XML (Yes)
	The standard should be written in characters that people can read (Lesk, 1995).	N/A
Complexity	Human readability, compression, variety of features (Rog & van Wijk, 2008; Wijk & Rog, 2007).	N/A
	Simple raster formats are preferred (Puglia et al., 2004).	N/A
Compression Algorithms	The format uses standard algorithms (Puglia et al., 2004).	N/A
Accessibility	To prohibit encryption in the file trailer (Sullivan, 2006)	PDF/A (Yes)
Component Reuse	Not defined (Sahu, 2006)	PDF (No) HTML (Limited) SGML (Excellent) XML (Excellent)
Repurposing	Not defined (Sahu, 1999)	PDF (Limited) HTML (Limited) SGML (Excellent) XML (Excellent)
Packaging formats	In general, packaging formats should be acceptable as transfer mechanisms for image file formats (Puglia et al., 2004).	Zip (Yes)
Significant Properties	The format accommodates high-bit, high-resolution (detail), color accuracy, and multiple compression options (Puglia et al., 2004).	N/A
Processability	The requirement to maintain a processable version of the record to have any reuse value (Brown, 2003)	Conversion of a word-processed document into PDF format. (No)
Searching	Not defined (Sahu, 2006)	PDF (Limited) HTML (Good) SGML (Excellent) XML (Excellent)
No Definite Term	To 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 XCL (Yes)
	To make transferring data easy (Johnson, 1999)	N/A XML (Yes)
	A format that is easy to restore and understand by both humans and machines (Müller et al., 2003)	N/A XML (Yes)

		Inability to be backed out into a usable format (Potter, 2006)	PDFs (No)
2. M E T A D A T A	Self-Documentation	Self-documenting digital objects that contain basic descriptive, technical, and other administrative metadata (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes) XML (Yes)
		Metadata and technical description of format embedded (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Limited) Microsoft Word (Limited)
		The ability of a digital format to hold (in a transparent form) metadata beyond that needed for basic rendering of the content (Arms & Fleischhauer, 2006)	N/A
	Self-Documenting	To contain its own description (Abrams et al., 2005)	N/A
	Documentation	Deep technical documentation publicly and fully is available. It is maintained for older versions of the format (Puglia et al., 2004).	N/A
	Metadata Support	File formats making provision for the inclusion of metadata (Brown, 2003)	TIFF (Yes) Microsoft Word 2000 (Yes)
		Not defined (Kenney, 2001)	FIFF 6.0 (Yes) GIF 89a (Yes) JPEG (Yes) Flashpix 1.0.2 (Yes) ImagePac, Photo CD (No) PNG 1.2 (Yes) PDF (Yes)
		Not defined (Sahu, 2006)	N/A
	Metadata	The format allows for self-documentation (Puglia et al., 2004).	N/A
	Content-Level Description	Not presentation-level description; structural markup, not formatting (Barnes, 2006)	PDF (No) DocBook (Yes) TEI (Yes) XHTML (Yes) XML (Yes)
	Content-Level, Not Presentation-Level, Descriptions	Where possible, the labeling of items should reflect their meaning, not their appearance (Lesk, 1995).	SGML (Yes)
	Self-Describing	Many different types of metadata are required to decipher the contents of a file (Folk & Barkstrom, 2003).	N/A
	Self-Describing Files	Embed metadata in PDF files (Sullivan, 2006)	PDF/A (Adobe Extensible Metadata Platform Required)
	Formal (BNF- or XML-Like) Description of Format	To create new readers solely on the basis of formal descriptions of the file content (Folk & Barkstrom, 2003)	N/A
	No Definite Term	Its self-describing tags identify what your content is all about (Johnson, 1999).	N/A XML (Yes)
A format for strong descriptive and administrative metadata and the complete content of the document (Müller et al., 2003)		N/A XML (Yes)	

3. O P E N N E S S	Disclosure	Authoritative specification publicly available (Abrams et al., 2005)	PDF/A (Yes) Microsoft Word (No)
		The degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content (CENDI, 2007; Hodge & Anderson, 2007; Arms & Fleischhauer, 2006)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes) XML (Yes)
		Authoritative specification is publicly available (Sullivan, 2006).	PDF/A (Yes)
	Open Availability	No proprietary formats (Barnes, 2006)	ODF (Yes) GIF (No) PDF (No) RTF (No) Microsoft Word (No)
		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)
	Openness	Standardization, restrictions on the interpretation of the file format, reader with freely available source (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Yes) MS Word (No)
		A standard is designed to be implemented by multiple providers and Guide 5: File Formats for Digital Masters employed by a large number of users (Frey, 2000).	N/A
		Formats that are described by publicly available specifications or open-source source code can, with some effort, be reconstructed later: (1) open publicly available specification, (2) specification in public domain, (3) viewer with freely available source, (4) viewer with GPL'ed source, (5) not encrypted (Clausen, 2004).	N/A
	Open-Source Software or Equivalent	To move toward obtaining open-source arrangements for all parts of the file format and associated libraries (Folk & Barkstrom, 2003)	N/A
	Open Standard	Formats for which the technical specification has been made available in the public domain (Brown, 2003)	JPEG (Yes) PDF (Limited) ASCII (Limited)
		Not defined (Sahu, 2006)	N/A
	Standard/ Proprietary	Not defined (Kenney, 2001)	FIFF 6.0 (Yes) GIF 89a (Yes) JPEG (Yes) Flashpix 1.0.2 (Yes) ImagePac, Photo CD (No) PNG 1.2 (Yes) PDF (Yes)
	Nonproprietary Formats	The specification is independent of a particular vendor (Public Records Office of Victoria, 2004).	N/A
	No Definite Term	To avoid vendor-lock (Potter, 2006)	ODF (Yes)

4. I N T E R O P E R A B I L I T Y	Interoperability	Is the format supported by many software applications/OS platforms or is it linked closely with a specific application (Puglia et al., 2004)?	N/A
		The ability to exchange electronic records with other users and IT systems (Brown, 2003)	N/A
		Not defined (Sahu, 2006)	N/A
	Data Interchange	Not defined (Sahu, 2006)	PDF (No) HTML (Limited) SGML (Excellent) XML (Excellent)
	Compatibility	Compatibility with prior versions of data set definitions often is needed for access and migration considerations (Frey, 2000).	N/A
	Stability	Compatibility between versions (Folk & Barkstrom, 2003)	N/A
		Stable, not subject to constant or major changes over time (Brown, 2003)	N/A
		The format is supported by current applications and backward compatible, and there are frequent updates to the format or the specification (Puglia et al., 2004).	N/A
		Not defined (Sahu, 2006).	N/A
	Scalability	The design should be applicable both to small and large data sets and to small and large hardware systems (Frey, 2000).	N/A
Markup Compatibility and Extensibility	To support a much broader range of applications (ECMA, 2008)	N/A XML (Yes)	
Suitability for a Variety of Storage Technologies	The format should not be geared toward any particular technology (Folk & Barkstrom, 2003).	N/A	
No Definite Term	To allow data to be shared across information systems and remain impervious to many proprietary software revisions (Potter, 2006)	OpenOffice (Yes)	
5. I N D E P E N D E N C E	Device Independencies	Can be reliably and consistently rendered without regard to the hardware/software platform (Abrams et al., 2005)	PDF/A (Yes) TIFF (No)
		Static visual appearance can be reliably and consistently rendered and printed without regard to the hardware or software platform used (Sullivan, 2006).	PDF/A (Yes) PDF/X (Yes)
		This is a very important aspect for master files because they will be most likely used on various systems (Frey, 2000).	N/A
	Independent Implementations	Independent implementations help ensure that vendors accurately implement the specification (Public Records Office of Victoria, 2004).	N/A
	External-Dependency	Degree to which the format is dependent on specific hardware, operating system, or software for rendering or use and the complexity of dealing with those dependencies in future technical environments (Arms & Fleischhauer, 2006)	N/A
	External Dependencies	The 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 (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Limited) PDF/A (No) TIFF_G4 (No) XML (No)

	Portability	A format that makes extensive use of specific hardware or operating system features is likely to be unusable when that hardware or operating system falls into disuse. A format that is defined in an independent way will be much easier to use in the future: (1) independent of hardware; (2) independent of operating system; (3) independent of other software; (4) independent of particular institutions, groups, or events; (5) widespread current use; (6) little built-in functionality; and (7) single version or well-defined versions (Clausen, 2004).	N/A
	Monitoring Obsolescence	Information gathered through regular web harvesting can give us some information about what file types are approaching obsolescence, at least for the more frequently used types (Clausen, 2004).	N/A
	No Definite Term	A human-readable text format and internationalized character sets are supported (Müller et al., 2003).	N/A XML (Yes)
		Not dependent on specific hardware, not dependent on specific operating systems, not dependent on one specific reader, not dependent on other external resources (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Limited) Microsoft Word (Little)
The format requires a plug-in for viewing if appropriate software is not available or relies on external programs to function (Puglia et al., 2004).		N/A	
6. P R E S E R V E N T I O N	Distributing Page Image	Not defined (Sahu, 2006)	PDF (Excellent) HTML (Good) SGML (Good) XML (Good)
	Normal Rendering	Not defined (CENDI, 2007; Hodge & Anderson, 2007).	PDF (Yes) PDF/A (Limited) TIFF_G4 (Yes) XML (Yes)
	Presentation	Preservation of its original look and feel (Brown, 2003)	N/A
	Self-Containment	Everything that is necessary to render or print a PDF/A file must be contained within the file (Sullivan, 2006).	PDF/A (Yes)
	Self-Contained	To contain all resources necessary for rendering (Abrams et al., 2005)	N/A
	Beyond Normal Rendering	Not defined (CENDI, 2007; Hodge & Anderson, 2007).	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes) XML (Limited)
7. A U T H E N T I C I T Y	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
	Provenance Traceability	Ability to trace the entire configuration of data production (Folk & Barkstrom, 2003)	N/A
	Integrity of Layout	Not defined (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (N/A) XML (Yes)
	Integrity of Rendering of Equations	Not defined (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (N/A) XML (Limited)
	Integrity of Structure	Not defined (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Limited) PDF/A (Limited) TIFF_G4 (N/A)

			XML (Yes)
8. A D O P T I O N	Adoption	Degree to which the format is already used by the primary creators, disseminators, or users of information resources (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Yes) PDF/A (Yes) TIFF_G4 (Yes) XML (Yes)
		Worldwide usage, usage in the cultural heritage sector as archival format (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Yes) Microsoft Word (Limited)
		The degree to which the format is already used by the primary creators, disseminators, or users of information resources (Arms & Fleischhauer, 2006)	N/A
		Widespread use may be the best deterrent against preservation risk (Abrams et al., 2005).	TIFF (Yes)
		The format is widely used by the imaging community in cultural institutions (Puglia et al., 2004).	N/A
		Flexibility of implementation to promote its wide adoption (Sullivan, 2006)	PDF/A (Yes)
	Popularity	A format that is widely used (Folk & Barkstrom, 2003)	N/A
	Widely Used Formats	It is far more likely that software will continue to be available to render the format (Public Records Office of Victoria, 2004).	N/A
	Ubiquity	Popular formats supported by as much software as possible (Brown, 2003)	N/A
		Not defined (Sahu, 2006)	N/A
Continuity	The file format is mature (Puglia et al., 2004)	N/A	
9. P R O T E C T I O N	Technical Protection Mechanism	Password protection, copy protection, digital signature, printing protection and content extraction protection (Rog & van Wijk, 2008; Wijk & Rog, 2007)	PDF/A-1 (Limited) Microsoft Word (Limited)
		Implementation of a mechanism such as encryption that prevents the preservation of content by a trusted repository (CENDI, 2007; Hodge & Anderson, 2007)	PDF (Yes) PDF/A (No) TIFF_G4 (No) XML (No)
		It must be able to replicate the content on new media, migrate and normalize it in the face of changing technology, and disseminate it to users at a resolution consistent with network bandwidth constraints (Arms & Fleischhauer, 2006).	N/A
		No encryption, passwords, etc. (Abrams et al. (2005)	N/A
	Protection	The format accommodates error detection, correction mechanisms, and encryption options (Puglia et al., 2004).	N/A
Source Verification	Cryptographic encoding of files or digital watermarks without overburdening the data centers or archives (Folk & Barkstrom, 2003)	N/A	

10. P R E S E R V A T I O N	Preservation	The format contains embedded objects (e.g., fonts, raster images) or links to external objects (Puglia et al., 2004).	N/A
	Long-Term Institutional Support	To ensure the long-term maintenance and support of a data format by placing responsibility for these operations on institutions (Folk & Barkstrom, 2003)	N/A
	Ease of Transformation/ Preservation	The format will be supported for fully functional preservation in a repository setting, or the format guarantee can currently only be made at the bitstream (content data) level (Puglia et al., 2004).	N/A
	No Definite Term	To create files with either a very high or very low preservation value (Becker et al., 2008a, Becker et al., 2008b)	PDF (No) TIFF (No)
11. R E F E R E N C E	Citability	A machine-independent ability to reference or “cite” the individual data element in a stable way (Folk & Barkstrom, 2003)	N/A
	Referential Extensibility	Ability to build annotations about new interpretations of the data (Folk & Barkstrom, 2003)	N/A
	No Definite Term	An open and established notation (Müller et al., 2003)	N/A XML (Yes)
		Data is easily repurposed via tags or translated to any medium (Johnson, 1999)	N/A XML (Yes)
Creating, using, and reusing tags is easy, making it highly extensible (Johnson, 1999).		N/A XML (Yes)	
12. O T H E R S	Transparency	Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor (CENDI, 2007, Hodge & Anderson, 2007).	PDF (Limited) PDF/A (Limited) TIFF_G4 (Limited) XML (Yes)
		In natural reading order (Sullivan, 2006).	PDF/A (Yes) Microsoft Notepad (Yes)
		The degree to which the format is already used by the primary creators, disseminators, or users of information resources (Arms & Fleischhauer, 2006)	N/A
		Amenable to direct analysis with basic tools (Abrams et al., 2005)	N/A
	Ample Comment Space	To allow rich metadata (Barnes, 2006)	N/A
		Items should be labeled, as far as possible, with enough information to serve for searching or cataloging (Lesk, 1995).	TIFF (Yes)
		A digital format may inhibit the ability of archival institutions to sustain content in that format (Arms & Fleischhauer, 2006).	N/A

Table Bibliography

Abrams, Stephen et al. 2005. "PDF-A: The Development of a Digital Preservation Standard." Paper presented at the 69th Annual Meeting for the Society of American Archivists, New Orleans, Louisiana, August 14–21, <http://www.aiim.org/documents/standards/PDF-A.ppt> (accessed November 21, 2011).

Arms, Caroline R. and Carl Fleischhauer. 2006. "Sustainability of Digital Formats: Planning for Library of Congress Collections." <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml> (accessed November 21, 2011).

Barnes, Ian. 2006. "Preservation of Word Processing Documents." http://apsr.anu.edu.au/publications/word_processing_preservation.pdf (accessed November 21, 2011).

Becker, Christoph et al. 2008. "A Generic XML Language for Characterising Objects to Support Digital Preservation." In *Proceedings of the 2008 ACM Symposium on Applied Computing*, Fortaleza, Ceara, Brazil, March 16–20.

Becker, Christoph et al. 2008. "Systematic Characterization of Objects in Digital Preservation: The eXtensible Characterization Language." *Journal of Universal Computer Science* 14, no 18: 2936–2952.

Brown, Adams. 2003. "The National Archives. Digital Preservation Guidance Note: Selecting File Formats for Long-Term Preservation." <http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf> (accessed November 21, 2011).

CENDI Digital Preservation Task Group. 2007. "Formats for Digital Preservation: A Review of Alternatives and Issues." http://www.cendi.gov/publications/CENDI_PresFormats_WhitePaper_03092007.pdf (accessed November 21, 2011).

Clausen, Lars R. 2004. "Handling File Formats." <http://netarchive.dk/publikationer/FileFormats-2004.pdf> (accessed November 21, 2011).

ECMA. 2008. "Office Open XML File Formats—Part 1." 2nd ed. <http://www.ecma-international.org/publications/standards/Ecma-376.htm> (accessed November 21, 2011).

Folk, Mike, and Bruce Barkstrom. 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, Houston, TX, May 27–31. http://www.hdfgroup.org/projects/nara/Sci_Formats_and_Archiving.pdf (accessed November 21, 2011).

-
- Frey, Franziska. 2000. "5. File Formats for Digital Masters." In *Guides to Quality in Visual Resource Imaging*, Research Libraries Group and Digital Library Federation.
<http://imagedigital.esteticas.unam.mx/PDF/Guides.pdf> (accessed November 21, 2011).
- Hodge, Gail and Nikkia Anderson. 2007. "Formats for Digital Preservation: A Review of Alternatives and Issues." *Information Services & Use* 27: 45–63.
- Johnson, Amy Helen. 1999. "XML Xtends its Reach: XML Finds Favor in Many IT Shops, but It's Still Not Right for Everyone." *Computerworld* 33, no. 42: 76–81.
- Lesk, Michael E. 1995. "Preserving Digital Objects: Recurrent Needs and Challenges." In *Proceedings of the 2nd NPO Conference on Multimedia Preservation*. Brisbane, Australia.
<http://www.lesk.com/mlesk/auspres/aus.html> (accessed November 21, 2011).
- Müller, Eva et al. 2003. "Using XML for Long-Term Preservation: Experiences from the DiVA Project." In *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations*. Berlin, May: 109–116, <https://edoc.hu-berlin.de/conferences/etd2003/hansson-peter/PDF/index.pdf>
(accessed December 8, 2012).
- Potter, John Michael. 2006. "Formats Conversion Technologies Set to Benefit Institutional Repositories."
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.7881\u0026rep=rep1\u0026type=pdf> (accessed November 21, 2011).
- Public Records Office of Victoria (Australia). 2006. "Advice on VERS Long-Term Preservation Formats PROS 99/007 (Version2) Specification 4." Department for Victorian Communities.
http://prov.vic.gov.au/wp-content/uploads/2012/01/VERS_Advice13.pdf (accessed November 21, 2011).
- Puglia, Steven, Jeffrey Reed, and Erin Rhodes. 2004. "Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files—Raster Images." US National Archives and Records Administration.
<http://www.archives.gov/preservation/technical/guidelines.pdf> (accessed November 21, 2011).
- Rog, Judith, and Caroline van Wijk. 2008. "Evaluating File Formats for Long-term Preservation." National Library of the Netherlands.
http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf (accessed November 21, 2011).
- Sahu, D.K. 2004. "Long Term Preservation: Which File Format to Use." Presentation at Workshops on Open Access & Institutional Repository, Chennai, India, May 2–8,
http://openmed.nic.in/1363/01/Long_term_preservation.pdf (accessed November 21, 2011).
- Sullivan, Susan J. 2006. "An Archival/Records Management Perspective on PDF/A." *Records Management Journal* 16, no. 1: 51–56.

van Wijk, Caroline, and Judith Rog. 2007. "Evaluating File Formats for Long-Term Preservation." Presentation at International Conference on Digital Preservation, Beijing, China, Oct 11-12. http://ipres.las.ac.cn/pdf/Caroline-iPRES2007-11-12oct_CW.pdf (accessed November 21, 2011).