InterPARES 2 Project
International Research on Permanent Authentic Records in Electronic Systems

Title:	**General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation**
Status:	Final (public)
Version:	1.1
Submission Date:	December 2006
Release Date:	March 2007
Author:	The InterPARES 2 Project
Writer(s):	Evelyn Peters McLellan, Insurance Corporation of British Columbia
Project Unit:	Domain 3
URL:	http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf [English] http://www.interpares.org/display_file.cfm?doc=ip2_file_formats_fichiers_numériques.pdf [French]

Table of Contents

Introduction **1**

1. Terminology **1**
- 1.1 What is a file format? 1
- 1.2 “Open” file formats 3
- 1.3 “Standard” file formats 4
- 1.4 “Stable” file formats 5
- 1.5 Standardizing terms 5

2. Selection criteria **6**
- 2.1 Widespread use 6
- 2.2 Non-proprietary origin 7
- 2.3 Availability of specifications 8
- 2.4 Platform independence (interoperability) 8
- 2.5 Compression 10
- 2.6 Discussion of criteria 11

3. Policy implications **13**

4. Recommendations for developing and implementing policies **16**

Appendix A: list of repositories reviewed 17
Appendix B: URLs of documents reviewed 18

Bibliography 20

Selecting Digital File Formats for Long-Term Preservation

Introduction

In recent years, it has become common practice for digital records repositories, including archives, to accept certain digital file formats for long-term preservation while rejecting others. The questions of whether archives should limit the number of file formats for preservation, and, if so, on what criteria selection of formats should be based, raise important theoretical and policy issues that need to be addressed both by researchers in the field of digital preservation and by practicing archivists managing digital repositories. This paper attempts to answer these questions by presenting an analysis of current issues and trends in the selection of file formats for preserving digital records, an analysis which, in keeping with the scope of research of InterPARES 2, focuses on the areas of e-government, the sciences and the arts. The paper offers a mainly qualitative review of documentation on the Web sites of twenty repositories and four multi-institutional collaborative groups which have well established ingest policies and/or procedures or guidelines for agencies transferring records.[1] All of these repositories or groups either state which formats are accepted into the repository or give guidance on what formats are best suited to long-term preservation. A list of these repositories and research groups is provided in an appendix to this paper.

1. Terminology

The documentation review was designed to elicit information on the criteria various institutions use to select digital file formats for long-term preservation. However, understanding and evaluating the criteria was hampered to some extent by the lack of a uniform terminology to describe format characteristics.

1.1 What is a file format?

Most institutions surveyed do not define *file format*. The term is typically understood to refer to the structure for organizing data within a file. InterPARES 2 defines file format in part as “the organization of data within files, usually designed to facilitate the storage, retrieval, processing, presentation, and/or transmission of the data by software.”[2] The PREMIS *Data Dictionary for Preservation Metadata* defines file format as “a specific, pre-established structure for the organization of a digital file or bitstream.”[3] This pre-established structure includes how the data are encoded, which is the way in which the bits are

[1] The research for this paper was conducted with the assistance of Tracey Krause and Yvonne Loiselle, graduate students at the School of Library, Archival and Information Studies at the University of British Columbia.

[2] “File format,” The InterPARES 2 Project Dictionary, InterPARES 2 Web site, http://www.interpares.org/ip2/display_file.cfm?doc=ip2_dictionary.pdf.

[3] Preservation Metadata: Implementation Strategies (PREMIS), *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*. United States: On-Line Computer Library Center and Research Libraries Group, May 2005, p. 237. http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

interpreted to produce text, images and sound. Some types of encoding are synonymous with specific file formats; for example, MP3 encoding is used to encode the MP3 audio file format. However, many formats can have different encodings: even a "plain text" file can be encoded as ASCII, EBCDIC or Unicode, all of which have numerous variants. Encoding can be problematic in audio and video file formats, because the optimal encoding for storage and transmission often involves compression (removing bits from digital files to reduce their size), which can hinder preservation efforts.[4] A study by the Arts and Humanities Data Service (AHDS) notes that for the widely-used WAVE format, for example, there are approximately 100 different compression encodings, which can mean that the file can be opened by some software but not others.[5] Even TIFF, a format often used for long-term preservation of images, can be encoded in different ways, some of which involve lossy compression (a compression technique which results in some data being permanently lost).

The encoding issue is further complicated by the fact that TIFF, WAVE, AVI and other common image and audiovisual formats are not file "formats" per se, but rather file "wrapper formats" (also called "container formats"), which are designed to combine multiple bitstreams into a single file. The AHDS study notes that, with respect to the WAVE wrapper format, "it is only when the package is opened that the detailed requirements – the most important being the method of compression – are brought to light. If the software codec [i.e. the compression program] is specified and available, the file will play. If not, there is a serious problem that will probably require investigative software to diagnose…."[6] The study recommends AVI, QuickTime and WMV wrappers as archival formats for audiovisual objects only "as long as audio and video bitstreams within the wrappers are uncompressed or use lossless compression…."[7]

A number of the institutions surveyed list XML as a recommended or accepted file format. The question arises however, as to whether this is a file format per se. XML is in fact a meta-language which allows the user to develop specific sets of tags, or markup languages, to specify the structure of certain types of documents. GML (Geography Markup Language) and XHTML (Extensible Hypertext Markup Language) are subtypes of XML. Neither of these can be considered file formats in the strict sense of the term, but instead are text files with XML-defined tags. XML is more generic than a markup language, which at least contains a specified set of tags designed to structure data in a file. XML in itself is not, therefore, a file format, but only enables the structuring and definition of file formats and wrapper formats. The implications of accepting XML formats are developed more fully later on in this paper.

[4] See Caroline R. Arms and Carl Fleischhauer, compilers, *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Washington, D.C.: Library of Congress, updated March 6, 2006. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml. For a useful introduction to compression, see *File Formats and Compression*. United Kingdom: Technical Advisory Service for Images, March 2005. http://www.tasi.ac.uk/advice/creating/fformat.html#ff3.

[5] Andrew Wilson et al., *Moving Images and Sound Archiving Study*. Arts and Humanities Data Service, United Kingdom: Final Draft, June 2006, p. 29. http://roda.iantt.pt/?q=en/system/files/Moving+Images+and+Sound+Archiving+Study1.doc. Definition of codec inserted by this author.

[6] Ibid.

[7] Ibid., p. 48.

Defining *file format* is thus not always a simple task, and institutions issuing guidelines on preservation formats or specifying which formats may be accepted into a digital repository need to be clear on whether the “formats” they accept are file formats, wrapper formats or tagged files. For wrapper formats, constituent bitstream encodings need to be specified, and for XML files the file format, encoding and XML schema or DTD should be specified. For any formats having more than one version, the version number should also be specified, since different versions of a format can have different encoding specifications and other distinct characteristics.

1.2 “Open” file formats

Many of the institutions surveyed use terms such as *open* file format to describe file formats that have any or all of a number of characteristics. Chief among these is that the specification is published and freely available and that the format is created by non-proprietary (or open-source) software. However, it may also mean that the format is free of patent or royalty fees or the possibility of such fees being applied in the future, and/or that it is widely adopted. In many cases the meaning of the term is not clear. For example, the National Archives of Australia accepts file formats that are “both open and well documented,”[8] indicating that having a freely available specification is a characteristic that is separate from the characteristic of being open, while the Research Libraries Group and Digital Library Federation refer to “open and non-proprietary formats,”[9] which suggests that the characteristics of being open and non-proprietary are separate. The Technical Advisory Service for Images in the UK writes that a format selected for digital preservation should “be an open standard file format - proprietary formats should not be used, as there is uncertainty about the ability to open the file in the future;”[10] there is no mention of openness referring to the existence of a published, freely available specification. The EU-US Working Group on Spoken-Word Audio Collections suggests that “there exists what might be called an ‘openness spectrum,’ with degrees of protection applied by industry developers....Very comprehensive and fully public documentation is available for some proprietary content formats, e.g., TIFF.”[11] The UK Data Archive distinguishes between proprietary formats, “available” formats (proprietary formats with freely available specifications) and “open” formats, which are formats “created by a co-operative group that are then made freely available to anybody to use without restriction.”[12] It should be noted that

[8] Simon Davis, *Recordkeeping Issues Forum: Digital Preservation Strategy*. Australia: National Archives of Australia, November 19, 2002, p. 4. http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf.

[9] Franziska Frey, *Guides to Quality in Visual Resource Imaging*: *5. File Formats for Digital Masters*. United States: Research Libraries Group and Digital Library Federation, 2000. Section 2.3, *Pros and Cons of Various Formats.* http://www.rlg.org/visguides/visguide5.html.

[10] Technical Advisory Service for Images, *Advice Paper: Choosing a File Format.* United Kingdom: Technical Advisory Services, May 2006. http://www.tasi.ac.uk/advice/creating/format.html.

[11] Steve Renals and Jerry Goldman et al., *EU-US Working Group on Spoken-Word Audio Collections, Final Report*. European Union and United States: EU-US Working Group on Spoken-Word Audio Collections, June 18, 2003, p. 33. http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf.

[12] UK Data Archive, *UK Data Archive Preservation Policy*. Colchester: University of Essex, version 2.0, September, 2005, p. 19.
http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf.

"open" formats, if they are defined as being non-proprietary and having freely available specifications, are not necessarily the same as formats produced by open source software. The latter term describes software for which the code is made freely available and can be modified. Open source software does not always produce non-proprietary formats: for example, some open source software can be used to create PDFs, a proprietary format.[13]

1.3 "Standard" file formats

A smaller number of institutions use the term *standard* to characterize accepted or recommended file formats. This is a broad term used to indicate that the format has a number of desirable features. The Netherlands Institute for Scientific Information Services, for example, defines "standard image file formats" as being formats that are widely adopted, have freely available specifications, are highly interoperable, incorporate no data compression and are capable of supporting preservation metadata.[14] Library and Archives Canada refers to "de facto standard" formats, which, though proprietary, are "recognized formats and file types that have become industry standards because of their ubiquitous use and support, and not because they have been formally approved by a standards organization."[15] The Public Records Office of Victoria, Australia, recommends that in the absence of a non-proprietary format with a published specification, an "industry standard" format be selected:

> The basis for using an 'industry standard' format is economics. If the industry standard product has the major share of marketplace, it is unlikely to be replaced easily or quickly. Consequently, records in this format will be accessible for a reasonable period. Further, if the industry standard is ever replaced, the replacement products are almost certain to read the proprietary format of the product of the product they replace.[16]

Other organizations use the term standard to indicate that the format has been accepted or promoted by national or international standards organizations such as ANSI or ISO,[17] a criterion which may be substituted for or complemented by other criteria. The DAVID project, for example, states that "standard file formats owe their status to (official) initiatives for standardising," but adds that

[13] PDF (Portable Document Format) is subject to a number of patents held by its creator, Adobe Systems Inc. However, the specification is publicly available, and Adobe Systems offers royalty-free license rights to developers of software that conform to the specification. In addition, the International Standards Organization recently approved PDF/Archival (PDF/A), a stripped-down version of PDF 1.4, for use as a long-term preservation format (ISO 19005-1:2005, published on October 1, 2005).

[14] Rene van Horik, *Image Formats: Practical Experiences*. Netherlands: Netherlands Institute for Scientific Information Services, 2004. Erpanet presentation, Vienna, May 2004, p. 22. http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Horik.pdf.

[15] David L. Brown and Mike Swan. *Guidelines for Computer File Types, Interchange Formats and Information Standards*. Ottawa: Library and Archives Canada (LAC), version 1.1, June 28, 2004, section 1.3, *Concept*. http://www.collectionscanada.ca/information-management/002/007002-3017-e.html.

[16] Public Records Office of Victoria, *Advice 13: Long-Term Preservation Formats*. North Melbourne: Public Records Office Victoria, September 2, 2004, section 6.3*, Industry standard formats*. http://www.prov.vic.gov.au/vers/standard/advice_13/.

[17] American National Standards Institute and International Standards Organization, respectively.

this status may also be derived from “their widespread use.”[18] The DAVID documentation does not discuss in detail what constitutes an official initiative to standardize a file format. The Berkeley Art Museum/Pacific Film Archive refers to the use of “national and international standards whenever feasible” and states that durability is enhanced in digital content that “is in Standard formats that are well documented by a wide community and not under the control of any one company.” Berkeley also calls these “neutral formats”.[19]

1.4 “Stable” file formats

Library and Archives Canada, the Florida Center for Library Automation (FCLA) Digital Archives, the National Archives of the United Kingdom and the UK Data Archive refer to format *stability*. A stable file format appears to be one that is both backwards compatible (i.e., compatible with previous versions of the format) and well-supported by the software industry. Library and Archives Canada writes that AVI has become a *de facto* standard, “but Microsoft has announced that it will soon drop support for the format. In the short-term, AVI files should be converted to a more stable format because its prospects for future support are not good.”[20] The National Archives of the United Kingdom writes that “it is desirable that the format specification should be stable and not subject to constant or major changes over time. New versions of the format should also be backwards compatible.”[21] The FCLA and UK Data Archive recommend the use of stable formats but do not define the term.[22]

1.5 Standardizing terms

The terms used to describe characteristics of digital file formats that improve the prospects for successful long-term preservation are thus not consistently defined across institutions, and are sometimes not defined at all. Standardizing the definitions for such terms as *open*, *standard* and *stable* would be a useful step in assisting preserving institutions to communicate effectively with their donors, their parent institutions and other digital repositories. The use of the terms *open* and *standard* appears to be particularly problematic, but it will become clear from the rest of this paper that there is considerable variability in the terminology used to describe many other preservation-friendly characteristics of file formats as well.

[18] DAVID Project (Archivering in Vlaamse Instellingen en Diensten, or Digital Archiving in Flemish Institutions and Administrations), *Digital Archiving, Guideline and Advice 4: Standards for Fileformats*. Antwerp, 2003, p. 1. http://www.expertisecentrumdavid.be/davidproject/teksten/guideline4.pdf.

[19] Richard Rinehart and Guenter Waibel, *Strategies for Digital Media Asset Management*. California: UC Berkeley Art Museum/Pacific Film Archive, April 26, 2001. Unpublished internal policy document.

[20] Brown and Swan, *Guidelines*, op. cit., section 3.3.2.1. *Audio Video Interleave (AVI)*.

[21] Adrian Brown, *Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation*. Surrey, UK: National Archives of the United Kingdom, June 19, 2003, p. 6. http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf.

[22] “As a general rule, use platform-independent, vendor-independent, non-proprietary, stable, open and well-supported formats.” Florida Center for Library Automation (FCLA), *Recommended Data Formats for Preservation Purposes in the FCLA Digital Archive*. Gainesville, Florida: Florida Center for Library Automation, June 2005, p. 2. http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf; “The UKDA prefers to store files in formats that are based on stable, open and available standards….”, UK Data Archive, *Preservation Policy*, op. cit., p. 23. http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf.

2. Selection criteria

The documentation review reveals a number of digital file format features that are generally considered to make the formats suitable for long-term preservation. The most prominent of these are that they be widely adopted, non-proprietary, well-documented (i.e., have a freely available specification), platform independent (or interoperable), and either uncompressed or compressed using a lossless technique only. These criteria are discussed in detail below.

2.1 Widespread use

Eighteen of the twenty-four institutions[23] reviewed cite widespread adoption as a criterion for selecting file formats for long-term digital preservation. The On-line Computer Library Center, for example, states that in assessing risk factors, digital repositories need to consider whether a format is "widely accepted or simply a niche format."[24] The Massachusetts Institute of Technology accepts (but does not guarantee full preservation of) Microsoft Word and PowerPoint, Lotus 1-2-3 and WordPerfect files because the formats "are so popular that third party migration tools will likely emerge to help with format migration."[25] The National Archives of the United Kingdom favours formats that are in widespread use because this leads to continued support for the format by the software industry:

> 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 a range of software as possible, are therefore to be preferred where possible.[26]

Library of Congress argues that "if a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions."[27] Widespread use as a key element in the development of industry or de facto standards is described above.

Determining what constitutes widespread use is a subjective exercise. Cornell University Library's documentation states that a format should have "wide adoption by large consortia and groups, to increase the chances for well-defined migration paths,"[28] while the Netherlands Institute for Scientific Information

[23] For the sake of simplicity, "institutions" includes the four collaborative research groups listed in Appendix A of this paper.

[24] Andreas Stanescu, "Assessing the Durability of Formats in a Digital Preservation Environment: the INFORM Methodology," *D-Lib Magazine* 10(11), November 2004. http://www.dlib.org/dlib/november04/stanescu/11stanescu.html.

[25] Massachusetts Institute of Technology (MIT), *General DSpace FAQ*. Cambridge, Massachusetts: MIT Libraries, undated. http://libraries.mit.edu/DSpace-mit/about/faq.html.

[26] Brown, *Digital Preservation Guidance Note 1*, op. cit., p. 6.

[27] Arms and Fleischhauer, *Sustainability of Digital Formats*, op cit.

[28] Anne R. Kenney et al., *Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy: Final Project Report*. Ithaca, New York: Cornell University Library, May 2001, p. 8. http://www.library.cornell.edu/imls/IMLS-CULfinalreport2.pdf.

Services writes that the format should have been "used by a large community during a considerable period of time."[29] Library and Archives Canada recommends MPEG-2 over MPEG-4 for moving images in part because of the former's "market acceptance and penetration", while the MPEG-4 standard "has yet to be adopted by many software developers and manufacturers."[30] Library of Congress, acknowledging that level of use is difficult to quantify, writes that

> Evidence of wide adoption of a digital format includes bundling of tools with personal computers, native support in Web browsers or market-leading content creation tools, including those intended for professional use, and the existence of many competing products for creation, manipulation, or rendering of digital objects in the format.[31]

Interestingly, it adds that "a format that has been reviewed by other archival institutions and accepted as a preferred or supported archival format also provides evidence of adoption."[32] For the Department of Architecture at the Art Institute of Chicago, adoption by archives is key: the Department recommends TIFF and PDF because they are "widely utilized for archival purposes."[33]

2.2 Non-proprietary origin

Seventeen of the institutions surveyed consider non-proprietary formats preferable to proprietary formats. Some of these institutions refer to these as *open* formats, while others, such as the On-line Computer Library Center and the Library of Congress, stipulate in specific terms that a format should be free of royalty, license or patent fees.[34] Many institutions make exceptions for certain formats based on considerations such as widespread use, availability of specifications or ability to convert the files to non-proprietary formats. The Public Records Office of Victoria recommends using non-proprietary formats if possible, but uses PDF as one of its preservation formats as part of its Victorian Electronic Records Strategy (VERS), in part because of the availability of specifications.[35] The Ohio Electronic Records Committee recommends non-proprietary digital image formats, adding that if an institution accepts a proprietary format, it should "provide a bridge to a non-proprietary digital image file format."[36] Similarly, the Digital Preservation Coalition states that where only proprietary formats are available, "a crucial factor will be the export formats supported to allow data to be

[29] Horik, *Image Formats*, op. cit., p. 22.

[30] Brown and Swan, *Guidelines*, op. cit., sections 3.3.1.1, Groupe d'experts *pour le codage d'images animées (MPEG-2)*, and 3.3.2.2, *MPEG-4*.

[31] Arms and Fleischhauer, *Sustainability of Digital Formats*, op cit.

[32] Ibid.

[33] Art Institute of Chicago, Department of Architecture. *Collecting, Archiving and Exhibiting Digital Design Data*. Chicago: Kristine Fallon Associates Inc., 2004, p. 35. http://www.artic.edu/aic/collections/dept_architecture/dddreport/0C.pdf.

[34] Stanescu, "Assessing the Durability of Formats," op. cit.; Arms and Fleischhauer, *Sustainability of Digital Formats*, op. cit.

[35] Public Records Office of Victoria, *Advice 13*, op. cit., section 6.2.2, *Published Formats* and section 7.2 *PDF (Portable Document Format)*.

[36] Ohio Electronic Records Committee, *Revised Digital Imaging Guidelines: Guidelines for State of Ohio Executive Agencies and Local Governments*. Ohio: Ohio Electronic Records Committee, June 26, 2003. http://www.ohiojunction.net/erc/imagingrevision/revisedimaging2003.html.

moved out of...these proprietary environments."[37] As mentioned above, MIT accepts Microsoft and other proprietary formats based on their widespread use, but cautions that it may not be able to provide full preservation for these formats. Library of Congress, which does not use the term non-proprietary but instead refers to the presence or absence of patents, argues that

> The impact of patents may not be significant enough in itself to warrant treatment as an independent factor....[Widespread] adoption of a format may be a good indicator that there will be no adverse effect on the ability of archival institutions to sustain access to the content through migration, dynamic generation of service copies, or other techniques.[38]

2.3 Availability of specifications

For seventeen of the institutions, availability of specifications appears to be an important consideration in selecting file formats for long-term preservation. Some institutions use the term "documentation" instead of "specifications," or refer to file formats that are "well-documented." Cornell University Library, for example, advocates the selection of formats that have "thorough, nonproprietary development and documentation."[39] Whether documentation and specifications are the same thing is not entirely clear, however. For example, the On-line Computer Library Center cites as a format risk factor "whether the source or specification can be independently inspected," and also "whether [a format] is complex or poorly documented."[40] It would appear from this that documentation and specifications are separate things.

Most of the institutions recommending formats that have published or freely available specifications link this requirement to the requirement that the format be non-proprietary. However, quite a few institutions are willing to recommend or accept proprietary formats as long as they have published specifications. The institutions which accept PDF, such as the California Digital Library and the Victorian Public Records Office, are examples of this tendency. The Library of Congress goes further, arguing that for proprietary formats with unpublished specifications, "in the future, deposit of full documentation in escrow with a trusted archive would provide some degree of disclosure to support the preservation of information in proprietary formats for which documentation is not publicly available;"[41] thus the lack of freely available, published specifications need not be an insuperable barrier to long-term preservation.

2.4 Platform independence (interoperability)

This criterion is listed, in various terms, by thirteen of the institutions. The

[37] Neil Beagrie and Maggie Jones, *Digital Preservation Coalition Handbook*. United Kingdom: Digital Preservation Coalition, updated August 2006, section 5.2*, File Format and Standards*. http://www.dpconline.org/graphics/handbook/.
[38] Arms and Fleischhauer, *Sustainability of Digital Formats*, op. cit.
[39] Anne R. Kenney et al., *Preserving Cornell's Digital Image Collections*, op. cit., p. 8. http://www.library.cornell.edu/imls/IMLS-CULfinalreport2.pdf.
[40] Stanescu, "Assessing the Durability of Formats," op. cit.
[41] Arms and Fleischhauer, *Sustainability of Digital Formats*, op. cit.

Netherlands Institute for Scientific Information Services, for example, states that "a wide range of systems has to support the format,"[42] while the State and University Library, Arhus, and the Royal Library, Copenhagen, recommend formats that are independent of hardware, operating systems and other software.[43] Cornell University Library refers to "cross-platform consistency" and "minimal hardware/software dependencies,"[44] while the National Archives of the United Kingdom refers to "interoperability," with the following explanation:

> 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. This feature also tends to support the long-term sustainability of data by facilitating the migration of the data from one technical environment to another.[45]

Library of Congress lists "external dependencies," being "the degree to which a particular format depends on particular hardware, operating system, or software for rendering or use," as one of its sustainability factors.[46] The Digital Image Archive of Medieval Music (DIAMM) simply states that it uses TIFF because it is currently considered "the most widely compatible and easily transferable file format for images."[47]

In recent years, XML has been widely touted as a mechanism to facilitate data exchange and interoperability. Many of the institutions surveyed indicate that XML-tagged formats are acceptable or recommended for long-term preservation, particularly those using well-developed, standardized grammars such as GML and XHTML, which are being used to facilitate the interoperability of geospatial data and web pages, respectively. However, XML is not a panacea for data exchange problems. XML documents must conform to Document Type Definitions (DTDs), which declare their attributes, elements and syntax. According to an assessment of open document formats conducted for the European Commission, the fact that software companies and even individual records creators create DTDs can result in XML formats that are proprietary and software-dependent: "Documents that obey to [sic] different XML based formats or DTDs are not necessarily compatible. Conversion between the two formats could even prove extremely difficult, or even impossible." The report adds that "the potential for creation of 'proprietary' XML document formats has many in the Open Source community concerned."[48] Most of the institutions surveyed for this paper require that XML-based formats be accompanied by their DTDs when they are deposited in a digital repository. The maintenance of multiple DTDs and the

[42] Horik, *Image Formats*, op. cit., p. 22.

[43] Lars R. Clausen, main author. *Handling File Formats*. Denmark: State and University Library, Arhus, and the Royal Library, Copenhagen, May 2004, p. 12. http://netarchive.dk/publikationer/FileFormats-2004.pdf.

[44] Kenney et al., *Preserving Cornell's Digital Image Collections*, op. cit., p. 8.

[45] Brown, *Digital Preservation Guidance Note 1*, op. cit., p. 7.

[46] Arms and Fleischhauer, *Sustainability of Digital Formats*, op. cit.

[47] *Archiving*, Digital Image Archive of Medieval Music (DIAMM) Web site, United Kingdom: Universities of Oxford and London, June 15, 2006. http://ww.diamm.ac.uk/content/description/archiving.html.

[48] Valoris, *Comparative Assessment of Open Documents Formats: Market Overview*. European Commission: Dec. 30, 2003, p. 17. http://europa.eu.int/idabc/servlets/Doc?id=17982.

potential for the development of proprietary XML-based formats are likely to be problem areas for digital repositories, however, and any policy on formats for selection need to take these issues into consideration.

2.5 Compression

It is difficult to quantify attitudes towards compression because the mandates and acquisition policies of the institutions surveyed vary considerably. The Digital Image Archive of Medieval Music, for example, is a mainly photographic image archives that accepts only uncompressed TIFF images. At the other end of the spectrum is Library and Archives Canada, which acquires a wide variety of text, still image, audio and video files, in both uncompressed and compressed formats. However, it is possible to generalize that there is a marked preference for either uncompressed or losslessly compressed files if possible: four institutions accept or recommend uncompressed files only, and ten prefer or recommend uncompressed files but accept files with lossless compression. Only five of the institutions explicitly accept files with lossy compression. Three institutions, the On-line Computer Library Center, the EU-US Working Group on Spoken-Word Audio Collections and the State and University Library, Arhus/Royal Library, Copenhagen, identify compression as risky or problematic, but do not expressly reject it.[49]

Of the institutions accepting or recommending compressed files, most stipulate that only certain types of files may be compressed, or that certain types of compression are acceptable while others are not. The US National Archives and Records Administration (NARA), for example, prefers uncompressed TIFF files but accepts those with lossless compression, and will accept JPEG files with lossy compression under certain conditions.[50] Cornell University Library and Research Libraries Group/Digital Library Federation accept files with "visually lossless" compression (compression which removes some data without affecting visual image quality).[51] Some of the institutions, such as the Ohio Electronic Records Committee and the Public Records Office of Victoria, specify which compression standards or techniques must be used.[52] NARA provides general requirements and recommends compression techniques that meet the requirements, stipulating that the transferring agency "identify the file compression method used (if applicable) and the compression level (e.g., medium, high) selected for the image(s)."[53] The Library of Congress recommends that digital repositories "accept content compressed using publicly disclosed and widely adopted algorithms that are either lossless or have a

[49] Stanescu, "Assessing the Durability of Formats," op. cit.; EU-US Working Group on Spoken-Word Audio Collections, *Final Report,* pp. 23, 34; Clausen, *Handling File Formats*, op. cit., p. 16.

[50] U.S. National Archives and Records Administration (NARA), *Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: Digital Photographic Records*. Maryland, National Archives and Records Administration: undated.
http://www.archives.gov/records-mgmt/initiatives/erm-guidance.html.

[51] Kenney et al., *Preserving Cornell's Digital Image Collections*, op. cit., p. 37; Frey, *File Formats for Digital Masters*, op. cit., section 3.4, *Compression.*

[52] Ohio Electronic Records Committee, *Guidelines for State of Ohio Executive Agencies*, op. cit.; Public Records Office of Victoria, *Advice 13*, op. cit., sections 7.3 *TIFF (Tagged Image File Format)* and 7.4 *JPEG*.

[53] NARA, *Expanding Acceptable Transfer Requirements*, op. cit.

degree of lossy compression that is acceptable to the creator, publisher, or primary user as a master version."[54]

2.6 Discussion of criteria

Widespread adoption, non-proprietary origin, published specifications, interoperability and lack of compression or lossless compression appear to be the most important considerations in selecting digital file formats for long-term preservation, however variable the terms used to describe them, and any preservation policies or strategies that distinguish between acceptable and unacceptable file formats should take them into account. However, there is tacit acknowledgement by many institutions that formats having all these attributes are not always available and that acceptable substitutes must often be found. We have seen, for example, that the fact that a format is proprietary can be considered mitigated by the existence of a freely available specification or by the format's degree of popularity and support by the software industry. In many cases, there is simply no available format that meets the desired criteria. Library and Archives Canada admits that it has "attempted to balance the requirements for quality, stability, potential longevity and industry acceptance" and that non-proprietary or de-facto standard formats have been selected only "where possible."[55] The Florida Center for Library Automation recommends platform-independent and non-proprietary formats "as a general rule."[56] In its advice to data depositors, the UK Data Archive specifies that "open formats should be used whenever possible. Available formats such as Microsoft RTF and Adobe PDF should be considered next. Proprietary formats should only be considered as a last resort."[57]

The inability to find such formats becomes more pronounced in the scientific and artistic fields. Library of Congress notes that interoperability is a particular concern with regard to preserving scientific data, since "scientific datasets built from sensor data may be useless without specialized software for analysis and visualization, software that may itself be very difficult to sustain, even with source code available."[58] Recognizing this problem, NASA and the Goddard Space Center have developed their Common Data Format, along with an XML markup language called the Common Data Format Markup Language (CDFML), for exchanging and preserving astronomical and other types of remote sensor data. Other initiatives that include standardized formats for scientific data include the Planetary Data System (PDS),[59] The Standard Archive Format for Europe (SAFE), and the development of a number of XML markup languages, including Geography Markup Language (GML), Extensible Scientific Interchange Language (XSIL), XML-Formatted Data Unit (XFDU), Chemical Markup

[54] Arms and Fleischhauer, *Sustainability Factors*, op. cit.

[55] Brown and Swan, *Guidelines*, op. cit., section 1.3, *Concept*.

[56] FCLA, *Recommended Data Formats*, op. cit., p. 2.

[57] UK Data Archive, *Preservation Policy*, op. cit., p. 20. The UK Data Archive defines "available standards" as proprietary formats with freely available specifications, and "proprietary standards" as "formats that are owned by a company and not generally made available." Ibid.

[58] Arms and Fleischhauer, *Sustainability of Digital Formats*, op. cit.

[59] See InterPARES 2 Case Study 8 final report, William Underwood, *Mars Global Surveyor Data Records in the Planetary Data System: A Case Study*, September 2005 (draft).

Language (CML), Materials Markup Language (MatML), and the Data Documentation Initiative (DDI) for the social and behavioural sciences. Many of these initiatives are in relatively early stages, however, and have not achieved widespread adoption in the scientific community.[60] Moreover, they have not necessarily been accepted as standard formats for transfer to digital repositories, mainly because few established repositories preserve scientific data.[61] Similar difficulties are encountered with respect to digital artistic works. For digital images, the most widely-accepted format is TIFF. The Research Libraries Group and Digital Library Federation recommend this proprietary format because "from the currently available formats TIFF is the one that can be considered most 'archival'."[62] For digital audio files, where compression encoding is a critical issue, the authors of the AHDS study conclude that

> In the best case, we would have raw unedited master source files or sound streams to preserve, but it is unlikely that such objects will be offered for preservation in most cases. … All other things being equal, the best object to preserve is the one using an encoding that retains the highest bit depth and sampling rate, and with the least amount of compression.[63]

Richard Rinehart of the UC Berkeley Art Museum and Pacific Film Archive notes that his institution establishes "a small set of default formats, and then deviate from that only when necessary." He adds that this strategy works fairly well when it comes to documentation about works of art, but not for the works of art themselves:

> The main deviation from standard formats is our born-digital art and film collections. These works of digital art come to us in a variety of formats.…The preservation formats for some of this digital art will be driven by the nature of the work. We plan to adopt a variable media strategy where we can in fact port digital art to new media as they evolve, but each art work is a separate case in which we have to obey the wishes of the artist and our contract with them.[64]

The e-government area shows greater promise when it comes to producing records that are in preservation-friendly formats, because of the size of the organizations involved and the commonalities in programs and services that tend to emerge over time. All governments have word-processing and database systems, for example, and many are also adopting standardized web-based

[60] One of the more established of these projects is the Data Documentation Initiative, which released the first public version of its DTD in 2001. The DTD has been used by the California Digital Library and other major repositories. See the DDI Web site at http://www.icpsr.umich.edu/DDI/. For a discussion of the development of format and other data standards in the sciences in general, see John Rumble, Jr. et al., *Developing and Using Standards for Data and Information in Science and Technology*. Tennessee: Information International Associates, 2006. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/025.pdf.

[61] An occasional exception to this is geospatial data, because GIS has broad applicability in the area of e-government. NARA and LAC, for example, both accept geospatial data and specify acceptable transfer formats for such data.

[62] Frey, *File Formats for Digital Masters*, op. cit., section 2.3 *Pros and Cons of Various Formats*.

[63] Wilson et al., *Moving Images and Sound Archiving Study*, op. cit., p. 54.

[64] E-mail to Yvonne Loiselle, 11 August 2005.

service platforms and GIS databases. The need to exchange data within and between government agencies is likely to be well supported in the future by XML-based formats such as the OpenDocument Format developed by the Organization for the Advancement of Structured Information Standards (OASIS).[65] Where proprietary systems are used, normalization (conversion to preservation-friendly formats) is at present more likely to be successfully implemented for e-mail, word-processing files and databases, records typically produced by governments, than for those created by highly specialized scientific or artistic endeavours.

3. Policy implications

This paper has thus far not addressed the question of whether an archival repository should limit the number of file formats it accepts for preservation. Of course, most of the institutions surveyed have either explicitly or implicitly stated that this is their policy. For third-party digital repositories functioning as businesses and accepting digital records for preservation on a contractual basis, adopting such a policy is entirely a matter of choice, based on market considerations. However, for other types of institutions, certain dilemmas arise. Should a government archives refuse to accession certain formats used by its parent institution? Does it have the authority to do so? Is it ethical to do so? For organizations preserving artistic and scientific records, will this kind of policy result in records being lost or creative or scientific processes being unduly affected by the preferences of digital repositories?

Several major repositories, including MIT, Florida Center for Library Automation, OCLC Digital Archive, California Digital Library and NARA, have attempted to address these issues by offering varying levels of preservation, depending on format. FCLA's digital archives policy guide states that

> Any file format can be deposited in the FDA [Florida Digital Archives]. However, only files in supported formats can receive full preservation services with the aim of ensuring the continued usability of the file. Files in unsupported formats will be preserved in their original (submitted) version only (bit-level preservation).[66]

MIT's policy is very similar:

> At MIT, for the time being, we acknowledge the fact that the formats in which faculty create their research material are not something we can

[65] See Organization for the Advancement of Structured Information Standards, *OASIS Open Document Format for Office Applications (OpenDocument) Technical Committee FAQ,* 2006, at http://www.oasis-open.org/committees/office/faq.php. The development of OpenDocument was driven largely by the needs of the business community, but since business and government activities often generate the same types of documents, the format is expected to have wide applicability to governments as well as to businesses.

[66] Florida Center for Library Automation, *FCLA Digital Archive (FDA) Policy Guide*. Gainesville, Florida: Florida Center for Library Automation, version 1.1 December, 2004, p. 1. http://ww.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide1_1.pdf. The *Guide* adds that, for logical objects comprising files "in both supported and unsupported formats, there is no guarantee that the logical object will remain usable as intended."

> predict or control…. Because of this we've defined three levels of preservation for a given format: supported, known, or unsupported. Supported formats will be functionally preserved using either format migration or emulation techniques. Known formats are those that we can't promise to preserve (e.g. proprietary or binary formats) but which are so popular that we believe third party migration tools will emerge to help with format migration. Finally, unsupported formats are those that we don't know enough about to do any sort of functional preservation. For all three levels we will do bit-level preservation so that "digital archaeologists" of the future will have the raw material to work with if the material proves to be worth that effort.[67]

California Digital Library "will only guarantee preservation of the original bitstream" for "new or unknown file formats."[68] NARA offers three levels of preservation, based mainly on format, entitled Basic Preservation and Access, Enhanced Preservation and Access and Optimal Preservation and Access.[69] Accepting file formats without a commitment to preserve them fully is problematic for a number of reasons. It could lead to a false sense of security on the part of record donors, who may feel safe donating their records to respected, well-established digital repositories and who will assume that the records will, somehow, be accessible in the future even if they are not in a currently preferred format. It may pose policy and ethical problems for government archives, which are mandated not only to acquire but to preserve and make accessible the records of their parent institutions. It is also a commitment to an unknown and possibly unsupported amount of work sometime in the future, with no guarantee of a successful outcome. It is, however, an attempt to meet the needs of diverse populations of depositors, and may be a necessary policy for institutions with broad acquisition mandates.

To return to the question of whether archival institutions should limit the number of file formats they accept, this paper suggests that the answer is a qualified yes, depending on institutional mandates and relationships with parent and donor organizations. In most cases the issue cannot be avoided, unless an archives is willing either to accept all formats and commit to preserving them all (which is unlikely to be successful) or, like the institutions mentioned above, to accept records for which there are no fully-developed preservation plans and rely on the development of tools to migrate and render them in the future. The quantity of documentation available on the subject of desirable file formats attests to the fact that managers of digital repositories have reached the conclusion that limiting the number of file formats for accession is often necessary. Does this necessarily mean that certain digital records will, from the moment of creation, be destined to be short-lived? This depends in part on the responsiveness of the repository to the needs of its parent institutions and

[67] MIT, *General DSpace FAQ*, op. cit.
[68] CDL Digital Library Services Advisory Group, *CDL Guidelines for Digital Objects*. California Digital Library: version 2.0, draft, November 2005, p. 5. http://www.cdlib.org/inside/diglib/guidelines/cdl_gdo_v2_draft.pdf.
[69] Electronic Records Archives Program Management Office, *Introduction to Preservation and Access Levels Concepts*. Maryland: National Archives and Records Administration December 5, 2003, pp. 7-8. http://www.archives.gov/era/pdf/preservation-and-access-levels.pdf.

donors. The fact that widespread use of a given format is considered important by so many of the institutions surveyed here suggests that the needs of archives and the needs of at least some creators coincide. Interoperability and backwards compatibility are also preservation-friendly characteristics that are as likely to serve the needs of records creators as they are the requirements of archives. This is likely to be the case particularly in the areas of e-government and the sciences, but less so in the arts, where the on-going individual creative process is less dependent on data exchange between institutions.

Archivists can choose to go further than limiting which formats their archives will accept: they can actively promote format standards for records creation. In this case, there are several considerations to take into account, relating in part to whether the activities that produce the records fall within the e-government, scientific or artistic spheres. For government archives, promoting formats for digital records creation can be viewed as an attempt to promote best practices for records creation in general. A government archives can, and arguably should, develop expertise relating to preservable file formats that its parent institutions can draw upon when considering the development of new e-government programs and services. Library and Archives Canada is one government archives that explicitly promotes certain formats for records creation, formats which are also suitable for transferring data and records between agencies of the federal government.[70] Government archives have always had the responsibility of promoting good recordkeeping by their parent institutions and ensuring that the records of government activities are preserved. Recommending specific file formats for records creation based on their suitability for long-term preservation can be an extension of this traditional function.

Much the same can be said for institutions preserving scientific records, since the requirements to exchange data and to be able to rely on preserved records to document research results are similar to the data exchange and accountability needs of governments. While digital repositories need not dictate the types of records to be created, they can nevertheless make clear statements about what file format characteristics are most likely to result in scientific records being preserved successfully. The types of records being created, the juridical relationship of the repository to the records creators, and the mandates of both the repositories and creators will all affect how the archives and creators interact with respect to selection of formats for records creation.

It is harder to make these statements with respect to the arts. In many cases where digital art is involved the process of writing code and developing file formats may be part of the artistic creation process in which an archives should not interfere. Nonetheless, archives can still develop expertise with respect to format creation and selection and communicate this expertise to the artistic community. This relationship may take the form of consultant (the archives) and client (the creator), or simply be an on-going dialogue between preservers and creators.[71] It is possible even to conceive of an archives becoming a link

[70] Brown and Swan, *Guidelines*, op. cit., section 1.1, *Purpose and Scope*.

[71] This is, to a certain extent, the approach of the Variable Media Network, which "pairs artists with museum and media consultants" to convert their analogue and/or digital artworks to new media or to make the works media-independent. See the project's Web site at http://variablemedia.net/.

between artistic groups, where one group has relied on the archives to preserve a work of art and another group draws upon this experience in developing its own artistic works.

4. Recommendations for developing and implementing policies

Determining the criteria to be used for selecting file formats for long-term preservation, and, if appropriate, recommending these formats for records creation, should be seen as positive steps in the development of an archives' capacity to preserve digital records. This paper concludes, therefore, with the following recommendations for developing and implementing policies on selecting digital file formats for long-term preservation:

1. Clarify terminology: determine what is meant by terms such as *open*, *standard*, *stable* and *well-documented*, and define the terms in policy documents.

2. Distinguish between file formats, wrapper (or container) formats, and tagged formats such as XML-tagged files, and ensure that version, encoding and other characteristics are understood and fully specified.

3. For XML files, require that the files be well-formed and valid and accompanied by the relevant DTDs or schemas.

4. Choose widely-used, non-proprietary, platform-independent formats with freely available specifications where possible.

5. Specify whether compressed files are acceptable, and if so, state the type of compression permitted. Where possible, choose lossless compression techniques that conform to accepted international standards.

6. If it is not feasible to choose formats with the characteristics listed in recommendation 4, choose formats that are being preserved at other digital repositories and collaborate with these other repositories to develop preservation plans for them.

7. Where possible, work with records creators to ensure that they use software programs that create records in formats that meet the criteria listed in recommendation 4.

Appendix A: List of repositories reviewed

The on-line documentation of the following digital repositories was reviewed:

- Art Institute of Chicago, Department of Architecture
- Arts and Humanities Data Service, United Kingdom
- California Digital Library
- Cornell University Library
- Digital Image Archive of Medieval Music (DIAMM), Universities of Oxford and London
- Florida Center for Library Automation
- Library and Archives Canada
- Library of Congress
- Massachusetts Institute of Technology
- National Archives of Australia
- National Archives of the United Kingdom
- U.S. National Archives and Records Administration
- Netherlands Institute for Scientific Information Services
- Ohio Electronic Records Committee
- On-line Computer Library Center, United States
- Public Records Office of Victoria
- State and University Library, Arhus, and the Royal Library, Copenhagen
- Technical Advisory Service for Images, United Kingdom
- UC Berkeley Art Museum/Pacific Film Archive
- UK Data Archive

The documentation of the following collaborative or non-institutional groups was also included because it contained specific recommendations for file format selection for preservation purposes:

- DAVID Project (Digitale Archivering in Vlaamse Instellingen en Diensten, or Digital Archiving in Flemish Institutions and Administrations)
- Digital Preservation Coalition, United Kingdom
- EU-US Working Group on Spoken-Word Audio Collections
- Research Libraries Group and Digital Library Federation

Appendix B: URLs of documents reviewed, listed by repository[*]

This is a list of the URLs visited to develop the quantitative analysis of file format criteria presented in section 2 of this paper. The full citations for the documents reviewed are provided in the Bibliography at the end of the paper.

Art Institute of Chicago, Department of Architecture
http://www.artic.edu/aic/collections/dept_architecture/dddreport/0C.pdf

Arts and Humanities Data Service, United Kingdom
http://ahds.ac.uk/depositing/deposit-formats.htm

California Digital Library
http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#reqformats
http://www.cdlib.org/inside/diglib/guidelines/cdl_gdo_v2_draft.pdf

Cornell University Library
http://www.library.cornell.edu/imls/IMLS-CULfinalreport2.pdf

Digital Image Archive of Medieval Music (DIAMM), Universities of Oxford and London
http://www.diamm.ac.uk/content/description/archiving.html
http://www.diamm.ac.uk/content/description/quality.html

Florida Center for Library Automation
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Library and Archives Canada
http://www.collectionscanada.ca/information-management/002/007002-3017-e.html

Library of Congress
http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

Massachusetts Institute of Technology
http://libraries.mit.edu/dspace-mit/about/faq.html

National Archives of Australia
http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf

National Archives of the United Kingdom
http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf

[*] All sites last viewed December 3, 2006.

Netherlands Institute for Scientific Information Services
http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Horik.pdf

Ohio Electronic Records Committee
http://ww.ohiojunction.net/erc/imagingrevision/revisedimaging2003.html

On-line Computer Library Center, United States
http://www.dlib.org/dlib/november04/stanescu/11stanescu.html
http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf

Public Records Office of Victoria
http://www.prov.vic.gov.au/vers/standard/advice_13/

State and University Library, Arhus, and the Royal Library, Copenhagen
http://netarchive.dk/publikationer/FileFormats-2004.pdf

Technical Advisory Service for Images, United Kingdom
http://www.tasi.ac.uk/advice/creating/pdf/format.pdf
http://www.tasi.ac.uk/advice/creating/fformat.html#ff3

UK Data Archive
http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf

U.S. National Archives and Records Administration
http://www.archives.gov/research/arc/digitizing-archival-materials.pdf
http://www.archives.gov/records-mgmt/initiatives/erm-guidance.html
http://www.archives.gov/about/regulations/part-1228/l.html

DAVID Project (Digitale Archivering in Vlaamse Instellingen en Diensten, or Digital Archiving in Flemish Institutions and Administrations)
http://www.expertisecentrumdavid.be/davidproject/teksten/guideline4.pdf

Digital Preservation Coalition, United Kingdom
http://ww.dpconline.org/graphics/handbook

EU-US Working Group on Spoken-Word Audio Collections
http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/spokenword/SpokenWord.pdf

Research Libraries Group and Digital Library Federation
http://www.rlg.org/legacy/visguides/

Bibliography*

Arms, Caroline R., and Carl Fleischhauer. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Washington, D.C.: Library of Congress, updated March 6, 2006. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

Art Institute of Chicago, Department of Architecture. *Collecting, Archiving and Exhibiting Digital Design Data*. Chicago: Kristine Fallon Associates Inc., 2004. http://www.artic.edu/aic/collections/dept_architecture/dddreport/0C.pdf.

Beagrie, Neil, and Maggie Jones. *Digital Preservation Coalition Handbook*. United Kingdom: Digital Preservation Coalition, updated August 2006. http://www.dpconline.org/graphics/handbook.

Brown, Adrian, *Digital Preservation Guidance Note 1: Selecting File Formats for Long-Term Preservation*. Surrey, UK: National Archives of the United Kingdom, June 19, 2003. http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf.

Brown, David L., and Mike Swan. *Guidelines for Computer File Types, Interchange Formats and Information Standards*. Ottawa: Library and Archives Canada, version 1.1, June 28, 2004. http://www.collectionscanada.ca/information-management/002/007002-3017-e.html.

CDL Digital Library Services Advisory Group. *CDL Guidelines for Digital Images*, chapter 3, *Requirements*, March 10, 2005. http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#reqformats.

____________. *CDL Guidelines for Digital Objects*. California Digital Library Advisory Group: version 2.0, draft, November 2005. http://www.cdlib.org/inside/diglib/guidelines/cdl_gdo_v2_draft.pdf.

Clausen, Lars R., main author. *Handling File Formats*. Denmark: State and University Library, Arhus, and the Royal Library, Copenhagen, May 2004. http://netarchive.dk/publikationer/FileFormats-2004.pdf.

Data Documentation Initiative, Data Documentation Initiative Alliance, 2006. http://www.icpsr.umich.edu/DDI/.

DAVID Project*, Digital Archiving, Guideline and Advice 4: Standards for Fileformats*. Antwerp, 2003. http://www.expertisecentrumdavid.be/davidproject/teksten/guideline4.pdf.

* All sites last viewed on December 3, 2006.

Davis, Simon. *Recordkeeping Issues Forum: Digital Preservation Strategy*. Australia: National Archives of Australia, November 19, 2002. http://www.naa.gov.au/recordkeeping/rkpubs/fora/02nov/digital_preservation.pdf.

Digital Image Archive of Medieval Music (DIAMM) Web site. United Kingdom: Universities of Oxford and London, June 15, 2006. http://www.diamm.ac.uk/content/description/archiving.html.

Florida Center for Library Automation. *FCLA Digital Archive (FDA) Policy Guide*. Gainesville, Florida: Florida Center for Library Automation, version 1.1 December 2004. http://www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide1_1.pdf.

____________. *Recommended Data Formats for Preservation Purposes in the FCLA Digital Archive*. Gainesville, Florida: Florida Center for Library Automation, June 2005. http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf.

Frey, Franziska. *Guides to Quality in Visual Resource Imaging*: *5. File Formats for Digital Masters*. United States: Research Libraries Group and Digital Library Federation, 2000. http://www.rlg.org/legacy/visguides/.

James, Hamish. *AHDS Deposit Formats.* Arts and Humanities Data Service, United Kingdom, August 22, 2006. http://ahds.ac.uk/depositing/deposit-formats.htm.

Kenney, Anne R., et al. *Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy: Final Project Report*. Ithaca, New York: Cornell University Library, May 2001. www.library.cornell.edu/imls/IMLS-CULfinalreport2.pdf.

Massachusetts Institute of Technology. *General DSpace FAQ*. Cambridge, Massachusetts: MIT Libraries, undated. http://libraries.mit.edu/dspace-mit/about/faq.html.

Ohio Electronic Records Committee, *Revised Digital Imaging Guidelines: Guidelines for State of Ohio Executive Agencies and Local Governments*. Ohio: Ohio Electronic Records Committee, June 26, 2003. www.ohiojunction.net/erc/imagingrevision/revisedimaging2003.html.

On-Line Computer Library Center, *OCLC Digital Archive Preservation Policy and Supporting Documentation*, Dublin, Ohio, August 8, 2006. http://www.oclc.org/support/documentation/digitalarchive/preservationpolicy.pdf.

Organization for the Advancement of Structured Information Standards. *OASIS Open Document Format for Office Applications (OpenDocument) Technical Committee FAQ*. 2006. http://www.oasis-open.org/committees/office/faq.php.

Preservation Metadata: Implementation Strategies (PREMIS). *Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group*. United States: On-Line Computer Library Center and Research Libraries Group, May 2005. http://www.oclc.org/research/projects/pmwg/premis-final.pdf.

Public Records Office of Victoria. *Advice 13: Long-Term Preservation Formats*. North Melbourne: Public Records Office Victoria, September 2, 2004. http://www.prov.vic.gov.au/vers/standard/advice_13/.

Puglia, Steve, Jeffrey Reed and Erin Rhodes, *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images*. U.S. National Archives and Records Administration, June 2004. http://www.archives.gov/research/arc/digitizing-archival-materials.pdf.

Renals, Steve, and Jerry Goldman, et al. *EU-US Working Group on Spoken-Word Audio Collections, Final Report*. European Union and United States: EU-US Working Group on Spoken-Word Audio Collections, June 18, 2003. http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf.

Rinehart, Richard, and Guenter Waibel. *Strategies for Digital Media Asset Management*. California: UC Berkeley Art Museum/Pacific Film Archive, April 26, 2001. Unpublished internal policy document.

Rumble, Jr., John, et al. *Developing and Using Standards for Data and Information in Science and Technology*. Tennessee: Information International Associates, 2006. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/025.pdf.

Stanescu, Andreas. "Assessing the Durability of Formats in a Digital Preservation Environment: the INFORM Methodology," *D-Lib Magazine* 10(11), November 2004. http://www.dlib.org/dlib/november04/stanescu/11stanescu.html.

Technical Advisory Service for Images. *Advice Paper: Choosing a File Format*. United Kingdom: Technical Advisory Services, May 2006. http://www.tasi.ac.uk/advice/creating/pdf/format.pdf.

____________, *File Formats and Compression*. United Kingdom: Technical Advisory Service for Images, March 2005. http://www.tasi.ac.uk/advice/creating/fformat.html#ff3.

UC Berkeley Art Museum/Pacific Film Archive, e-mail from Richard Rinehart to Yvonne Loiselle, August 11, 2005.

UK Data Archive. *UK Data Archive Preservation Policy*. Colchester: University of Essex, version 2.0, September 2005. http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf.

U.S. National Archives and Records Administration. *NARA Electronic Records Management (ERM) Guidance on the Web*. Maryland, National Archives and Records Administration, various dates. http://www.archives.gov/records-mgmt/initiatives/erm-guidance.html. Various sections starting with the title *Expanding Acceptable Transfer Requirements….*

______________, *Regulations, Sub-chapter B, Records Management, Part 1228 -- Disposition of Federal Records, Subpart L* -- Transfer of Records to the National Archives of the United States. http://www.archives.gov/about/regulations/part-1228/l.html.

______________, Electronic Records Archives Program Management Office. *Introduction to Preservation and Access Levels Concepts*. Maryland: National Archives and Records Administration, December 5, 2003. http://www.archives.gov/era/pdf/preservation-and-access-levels.pdf.

Valoris*. Comparative Assessment of Open Documents Formats: Market Overview*. European Commission, December 30, 2003. http://europa.eu.int/idabc/servlets/Doc?id=17982.

van Horik, Rene. *Image Formats: Practical Experiences*. Netherlands: Netherlands Institute for Scientific Information Services, 2004. Erpanet presentation, Vienna, May 2004. http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Horik.pdf.

Variable Media Network, Guggenheim Museum and Daniel Langlois Foundation for Art, Science, and Technology. Web site, undated. http://variablemedia.net/.

Wilson, Andrew, et al., *Moving Images and Sound Archiving Study*. Arts and Humanities Data Service, United Kingdom: Final Draft, June 2006. http://roda.iantt.pt/?q=en/system/files/Moving+Images+and+Sound+Archiving+Study1.doc.