Format Policies

SMARTech Help

SMARTech

SMARTech, or Scholarly Materials And Research @ Georgia Tech, is a repository for the capture of the intellectual output of the Institute in support of its teaching and research missions. SMARTech connects stockpiles of digital materials currently in existence throughout campus to create a cohesive, useful, sustainable repository available to Georgia Tech and the world.

See the Mission and Collection Policy .

Why should I participate?

- Access barriers disappear
- Enhanced visibility, use, reputation
- Wide and rapid dissemination of intellectual output
- Supports classroom teaching
- Aids multidisciplinary inquiry
- Valuable recruiting tool
- Preservation and management of information assets
- Reduces duplication of effort
- Stimulates serendipitous discovery and collaboration

What types of materials can I submit and find in SMARTech?

SMARTech houses Georgia Tech research in digital format, including

- Annual Reports
- Conference Papers
- Electronic Theses & Dissertations
- Learning Objects
- Newsletters
- Pre-Prints/Post-Prints
- Proceedings
- Research Reports
- Simulations
- Technical Reports
- Web Pages
- White papers
- Working Papers

What file formats are accepted?

We accept standard formats that we can make a commitment to migrate and provide access to over the long term including:

Type	Description	File extension	Support level
Text/Images	Adobe PDF	pdf	supported
Text	HTML	htm, html	supported
Text	Rich Text Format	rtf	supported
Text	Text	txt	supported
Text	XML	xml	supported
Text	Microsoft Word	doc	known
Text	WordPerfect	wpd	known
Text	SGML	sgm, sgml	known

SMARTech Help

Images	JPEG	jpg, jpeg	supported
Images	GIF	gif	supported
Images	PNG	png	supported
Images	TIFF	tif, tiff	supported
Images	Post Script	ps, eps, ai	supported
Images	BMP	bmp	known
Images	Adobe Photoshop	pdd, psd	known
Images	Microsoft Powerpoint	ppt	known
Images	Photo CD	pcd	known
Video	MPEG	mpg, mpeg, mpe	supported
Video	Video Quicktime	mov, qt	known
Audio	WAV	wav	supported
Audio	MPEG	mpa, abs, mpeg	supported
Audio	AIFF	aiff, aif, aifc	supported
Audio	RealAudio	ra, ram	known
Audio	Basic	au, snd	known
Special	Microsoft Excel	xls	known
Special	Microsoft Project	mpp, mpx, mpd	known
Special	Microsoft Visio	vsd	known
Special	FileMaker/FMP3	fm	known
Special	LateX	latex	known
Special	Mathematica	ma	known
Special	Tex	tex	known
Special	TeXdvi	dvi	known

supported Items in this category can be used in the future through migration or emulation and the Library makes a commitment to do so.

known This category indicates that the specifics of the program code for that format are not public but the format is so widely used that the ability to use it in the future is almost certain.

How are materials in SMARTech preserved?

SMARTech is part of the MetaArchive Cooperative distributed digital preservation network. Georgia Tech Library participates in the MetaArchive program, an international effort for the preservation of electronic scholarly materials through the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP).

How do I start contributing to SMARTech?

- **email:** smartech@library.gatech.edu

Bentley Historical Library
Digital Curation Services
1150 Beal Avenue
Ann Arbor, MI 48109

Sustainable Formats and Conversion Strategies at the Bentley Historical Library

November 9, 2011
Version 1.0

Executive Summary

The Bentley Historical Library is committed to the long-term preservation of and access to its digital collections. Because the library must contend with thousands of potential file formats, Digital Curation Services has adopted a three-tier approach to facilitate the preservation and conversion of digital content:

- Tier 1: Materials produced in sustainable formats will be maintained in their original version.
- Tier 2: Common "at-risk" formats will be converted to preservation-quality file types to retain important features and functionalities.
- Tier 3: All other content will receive basic bit-level preservation.

This document provides further information on the Bentley Historical Library's accepted preservation formats and conversion strategies.

Please see the chart on pp. 3-5 for a list of sustainable preservation formats and at-risk formats that will be subject to conversion.

Tier 1: Preservation of Sustainable Formats

The library has identified a number of sustainable file formats (pp. 3-5) that are widely used and/or nonproprietary, many of which have been recognized as international standards by bodies such as the International Standards Organization (ISO), ECMA International, and the Organization for the Advancement of Structured Information Standards (OASIS). The longevity of these formats has furthermore been acknowledged by various peer institutions and experts in the digital curation community, including the Library of Congress's National Digital Information Infrastructure and Preservation Program.

Digital materials stored in these file formats should remain usable to researchers and administrative units at the University of Michigan for the foreseeable future and beyond. The Bentley Historical Library will therefore preserve the original version of content stored in these sustainable formats at the time of accession. Digital Curation Services will monitor community best practices and technological advances in case a migration to alternative preservation formats should prove necessary.

Visit http://fileinfo.com to find basic descriptions of file formats or search the PRONOM Technical Registry for format specifications and more in-depth information.

Tier 2: Conversion of At-Risk Formats

The digital curation community has long acknowledged the disadvantages posed by proprietary formats (for which only specific software may be used) and content encoded with "lossy" compression (i.e. compression that reduces the quality of the data to conserve space). The Bentley Historical Library will therefore convert the most common at-risk formats to preservation-quality sustainable formats. To ensure the authenticity of materials, the original version will be maintained alongside the preservation copy.

See pp. 3-5 for a list of at-risk formats and preservation targets; these strategies reflect the policies and practices of peer institutions as well as the National Digital Information Infrastructure and Preservation Program. Visit the Library of Congress "Sustainability of Digital Formats" site (http://www.digitalpreservation.gov/formats/index.shtml) for more information on preservation issues and descriptions of preferred formats.

Tier 3: Bit-Level Preservation of All Other Formats

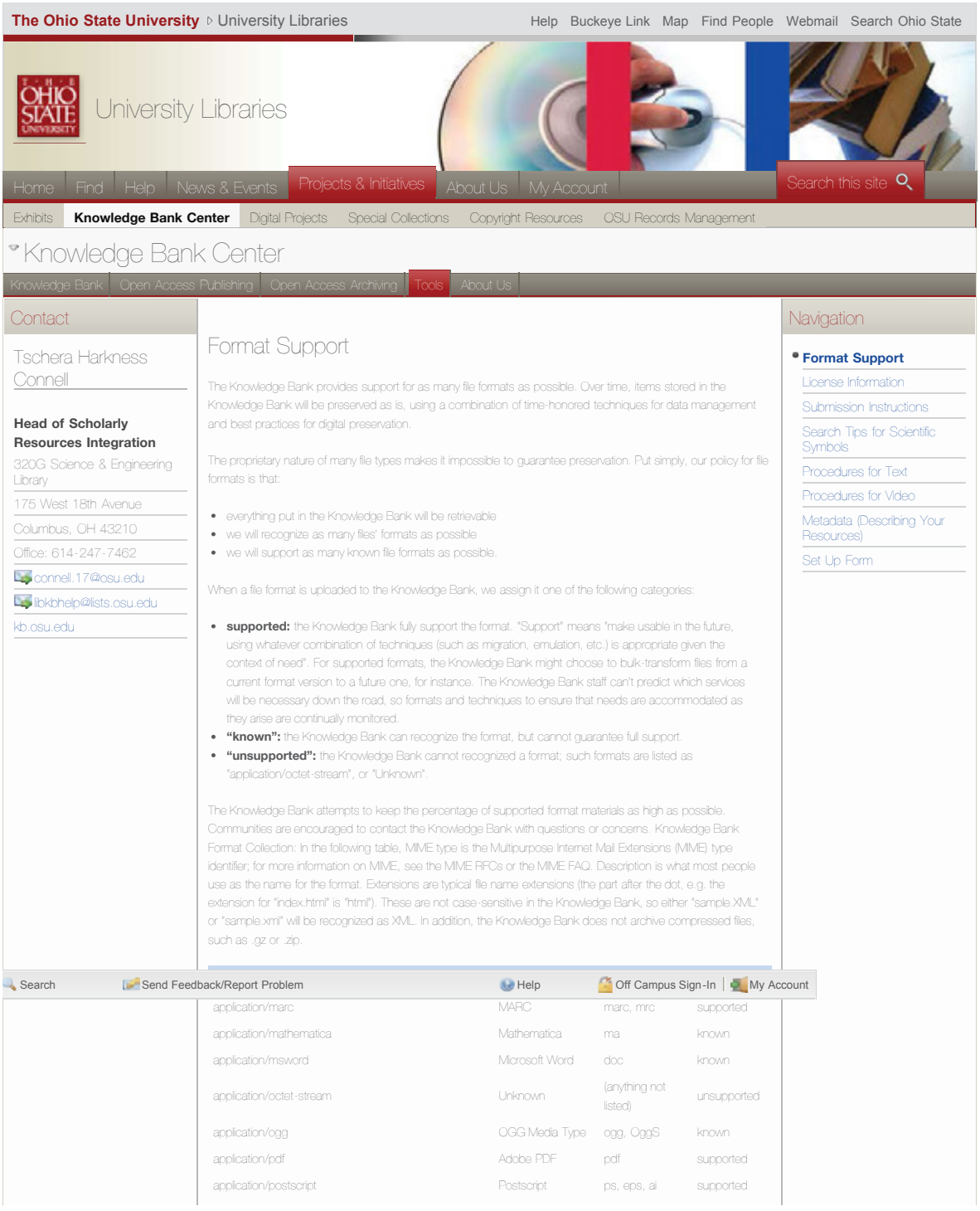
Because it is infeasible to create conversion plans for the tens of thousands of formats in existence, the Bentley Historical Library will ensure that digital holdings in other formats (i.e. ones not specifically identified in this document) will receive bit-level preservation. The use of integrity checks and regular replacement of storage media (conducted by trusted partners in the University of Michigan Library Information Technology division and Information and Technology Services) will preserve the raw data stored in these files (i.e. the "stream" of 0s and 1s) in its original state. The library concedes that hardware or software obsolescence may reduce the functionality of these files or render them inaccessible. At the same time, the faithful preservation of the content at the bit-level will allow the library to take advantage of future developments in emulation technology.

Tier 1: Preservation of Sustainable Formats	**Tier 2: Conversion Strategies for At-Risk Formats**	**Tier 3: Bit-Level Preservation**
Raster Images		
• TIFF: Tagged Image Format File • JPEG/JFIF: Joint Photographic Experts Group JPEG Interchange Format File (lossy compression) • JPEG 2000: Joint Photographic Experts Group (lossless compression) • GIF: Graphic Interchange Format • PNG: Portable Network Graphic	Convert the following to TIFF: • BMP: Windows Bitmap • PSD: Adobe Photoshop Document • RAW: Raw Image Data File • FPX: FlashPix Bitmap • PCD: Kodak Photo CD Image • PCT: Apple Picture File • TGA: Targa Graphic	All others
Vector Images		
• SVG: Scalable Vector Graphics File	Convert the following to SVG: • AI: Adobe Illustrator • WMF: Windows Metafile PS: Convert the following to PDF: • PS: PostScript • EPS: Encapsulated PostScript	All others
Audio Files		
• MIDI: Musical Instrument Digital Interface File • XMF: Extensible Music File • WAV: Waveform Audio File Format • AIFF: Audio Interchange File Format • MP3: Moving Picture Experts Group Layer 3 compression • OGG: Ogg Vorbis Audio File • FLAC: Free Lossless Audio Codec File	Convert the following to WAV: • WMA: Windows Media Audio • RA/RM: Real Audio • SND: Apple Sound File • AU: Sun Audio File	All others

Tier 1: Preservation of Sustainable Formats	**Tier 2: Conversion Strategies for At-Risk Formats**	**Tier 3: Bit-Level Preservation**
Video Files		
• MPEG-1/2: Moving Picture Experts Group • AVI: Audio Video Interleave File (uncompressed) • MOV: QuickTime Movie (uncompressed) • MP4: Moving Picture Experts Group (with H.264 encoding) • MJ2: Motion JPEG 2000 • MXF: Material Exchange Format File (uncompressed) • DV: Digital Video File (non-proprietary)	Convert the following to MP4 (with H.264 encoding): • SWF: Shockwave Flash • FLV: Flash Video • WMV: Windows Media Video • RV/RM: Real Video	All others
Office Documents and Text Files		
• DOCX: MS Word Open XML Document • XLSX: MS Excel Open XML Document • PPTX: PowerPoint Open XML Document • PDF: Portable Document Format • PDF/A: Portable Document Format (Archival) • TXT: Plain Text File • RTF: Rich Text Format File • XML: Extensible Markup Language Data File • CSV: Comma Separated Values File • TSV: Tab Separated Values File	Convert the following to Office Open XML: • DOC: MS Word Document • XLS: MS Excel Document • PPT: PowerPoint Document	All others
Email		
• MBOX: Mailbox File	Convert the following to MBOX: • EML: Email Message • PST: Outlook Personal Information Store File • Eudora mail, etc. (40 total)	All others

Tier 1: Preservation of Sustainable Formats	**Tier 2: Conversion Strategies for At-Risk Formats**	**Tier 3: Bit-Level Preservation**
Databases		
• SIARD: Software Independent Archiving of Relational Databases (open XML format) • CSV: Comma Separated Values File • MySQL SQL: Structured Query Language file; MySQL is an open source relational database management system	Convert the following into SIARD: • ACCDB or MDB: MS Access • SQL Server • Oracle Database	All Others

OHIO STATE UNIVERSITY
Format Support
http://library.osu.edu/projects-initiatives/knowledge-bank/tools/format-support/



OHIO STATE UNIVERSITY
Format Support
http://library.osu.edu/projects-initiatives/knowledge-bank/tools/format-support/

application/sgml	SGML	sgm, sgml	known
application/vnd.ms-excel	Microsoft Excel	xls	known
application/vnd.ms-powerpint	Microsoft Powerpoint	ppt	known
application/vnd.ms-project	Microsoft Project	mpp, mpx, mpd	known
application/vnd.openxmlformats-officedocument.presentationml.presentation	Microsoft PowerPoint XML	pptx	known
application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	Microsoft Excel XML	xlsx	known
application/vnd.openxmlformats-officedocument.wordprocessingml.document	Microsoft Word XML	docx	known
application/vnd.visio	Microsoft Visio	vsd	known
application/wordperfect5.1	WordPerfect	wpd	known
application/x-dvi	TeXdvi	dvi	known
application/x-filemaker	FMP3	fm	known
application/x-latex	LateX	latex	known
application/x-photoshop	Photoshop	psd, pdd	known
application/x-tex	TeX	tex	known
audio/x-aiff	AIFF	aiff, aif, aifc	supported
audio/basic	audio/basic	au, snd	known
audio/x-mpeg	MPEG Audio	mpa, abs, mpeg, mp3	known
audio/x-pn-realaudio	RealAudio	ra, ram	known
audio/x-wav	WAV	wav	known
image/gif	GIF	gif	supported
image/jpeg	JPEG	jpeg, jpg	supported
image/png	PNG	png	supported
image/tiff	TIFF	tiff, tif	supported
image/x-ms-bmp	BMP	bmp	known
image/x-photo-cd	Photo CD	pcd	known
text/comma-separated	CSV	csv	supported
text/css	CSS File	css	known
text/html	HTML	html, htm	supported
text/plain	Text	txt, asc	supported
text/richtext	Rich Text Format	rtf	supported
text/xml	XML	xml	supported
video/mpeg	MPEG	mpeg, mpg, mpe	known
video/quicktime	Video Quicktime	mov, qt	known

RUcore
Rutgers Community Repository

Archival standards for born-digital documents:
Recommended methods for keeping
stable preservation copies

Overview

As part of our plans to preserve student theses, dissertations, and newer editions of faculty texts and other culturally/academically significant documents, we inevitably will be tasked with preserving an increasing number of documents that originated electronically. These types of documents have been authored using various types of word processing and digital publishing software for decades, but the common practice had continued to be to print the final copy, and refer to the paper form as the final, finished product; the master original. Consequently, digital preservation would consist of scanning these analog objects back into a digital form, preserved electronically as scanned surrogates. Until very recently, we envisioned that scanning and digitizing from analog would comprise the bulk of how we digitally preserved all of our documents.

However, the increasing use of web-based publishing, online journals, and essentially paperless production has highlighted the benefits of seeking out the born-digital masters of preservation-worthy items whenever possible. Doing this affords us some advantages; namely, we can store the original in its most efficient digital form, often requiring less overhead and disk space while doing away with the quality challenges associated with scanning.

On the other hand, born digital preservation brings with it new challenges. Development of preservation standards for analog objects proved to be relatively simple, as the imaging industry laid much of the groundwork for us in terms of standardization across platforms. Further, development of future standards for digitized images, sound and video continues in an organized and orderly fashion, giving us plenty of time to contemplate migration to newer and better preservation formats.

Unfortunately, the same cannot be said for born digital documents. File formats for such objects vary widely, and the responsibility is upon us to identify a uniform set of file formats that we can adopt for preservation purposes.

As a result, a strategy for born digital document preservation must be adopted and followed that accomplishes the following:

- **Accurately renders** the formatting and content of the document, as intended by the creator of the document
- **Maintains stability** of the file format as well as possible. This may involve converting the document to archival formats, and storing both the original and the converted surrogate file.

Proposed Preservation Format Strategy: Multiple standards in play

Historically, born digital documents have been authored using a variety of different software packages, each with their own proprietary file formats. Early on, programs such as Wordstar, Wordperfect, Microsoft Works, ClarisWorks/AppleWorks, Adobe PageMaker, Quark Express, and others were distributed throughout the electronic document landscape.

More recently over the past decade, Microsoft Office has emerged as a de facto standard for general usage, with most businesses using it to create and distribute common document types. This usage has

resulted in a trickle-down effect to the consumer level on home computers and in academia as well. MS Office isn't perfect, however. The file formats used by Microsoft have evolved over the years as new versions have been released, and inconsistencies exist between versions in how document formatting is rendered.

At present, there are a number of formats developed by various consortia that attempt to solve the problem of maintaining a persistent document standard, and Microsoft itself has sought to modernize and make their document formats a formally accepted industry standard. Some of the more prevalent solutions include:

- **OpenXML:** A standard developed and endorsed by Microsoft and a consortium of other commercial software vendors, and is the standard document format used in the Microsoft Office suite beginning with Office 2007. These documents are often recognizable by their .docx, xlsx, and .pptx extensions.

- **OASIS OpenDocument (ODF):** An existing, open standard for file formats in use primarily in open source and "non-Microsoft" environments. These file formats are the default for OpenOffice.org and similar Free Software alternatives.

- **Portable Document Format/Archival (PDF and PDF/A):** A well-established standard with roots in Adobe PDF, a subset of which is now an ISO standard and a Library of Congress recognized format for digital document preservation.

There is also significant prevalence of legacy standards, a majority of which consists of legacy MS Office document types (.doc, .xls, .ppt, etc.) as well as more complex file formats for more intricate or specialized document types (LaTeX, Adobe InDesign, Illustrator, etc.). And finally, there are a multitude of document authoring platforms that are currently supported but have smaller market shares, such as Apple's iWork, current versions of Corel WordPerfect

Our choice of standards are based the ability to endure as technological advances continue to develop, and a widespread acceptance is key to ensuring easy migrating to newer standards when the time comes to retire existing choices.

The Recommendation: Our best case to preserve born digital documents while retaining longevity

Considering the state of the born digital document landscape as outline above, it is thus advisable that more than one preservation datastream for born-digital objects is utilized when possible. This strategy permits us to build redundancy into our repository, and ensure that regardless of whether one standard "wins out" over the other, our objects will remain with at least one relevant archival datastream. With that in mind, our strategy can be outlined as follows:

1. **Store the original document in its native format** when possible.
   In most cases, this will be an MS Office document, or a file from a similarly well-known software package. In some instances, the document we receive may already be rendered as a PDF file, in which case Step 2 below may not be necessary.

2. **Store an additional surrogate master in the form of a PDF/Archival file.**
   Most modern document authoring software, including MS Office and OpenOffice.org, have a

built-in capability to accurately "export" a document into a PDF version. This capability should be used when available to generate a faithful PDF file. Otherwise, the PDF/A can be generated using software available on RUcore platform.

Why PDF/A: An established standard to augment object datastreams

Although Portable Document Format has its roots in a proprietary system, recent efforts have proven fruitful – mainly thanks to Adobe, the creator of the file format – to have it recognized as an archival standard. PDF/A is defined by ISO 19005-1:2005, an ISO Standard that was published on October 1, 2005. According to the Library of Congress: "PDF/A is suggested as a preferred format for page-oriented textual (or primarily textual) documents when layout and visual characteristics are more significant than logical structure."[1]

The openness of this format has permitted a widening selection of software solutions to create archival PDFs from most digital documents. As indicated earlier, PDF "export" capability now exists on the market leading packages. Additionally, some computing platforms, namely OS X for Apple Mac computers and Linux environments, have a similar "print to PDF" feature standard as part of the operating system. Finally, free viewers exist for desktop and mobile computing platforms. This heavy documentation and wide accessibility make PDF/A a natural choice for acting as platform-independent method for preserving and making accessible born digital documents, without requiring users to purchase expensive, proprietary software to view the content.

Review provisions for special cases

The diversity that exists among born digital document formats virtually guarantees that a single standard will not address all use cases. In particular, this standard will not be well-suited to born digital documents that are formatted in such a way that a page-based presentation approach would be detrimental. In such a case, a review of how these documents were constructed will have to be undertaken, and the Digital Data Curator will need to consult the Cyber Infrastructure Working Group (CISC) and related subgroups on the best way to proceed.

[1] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml

RUcore
Rutgers Community Repository

Born Digital Still Images (Digital Photos):
Recommended Minimum Standards For Archival and Presentation Datastreams

(Note: This document addresses standards for born-digital still images only. For standards and requirements pertaining to digitization, i.e. the scanning of paper, slides or other analog media into digital images, please refer to the RUcore Digital Surrogate Guidelines.)

Introduction and Rationale

Since the inception of RUcore, a significant shift in the field of photography has taken place, as amateurs and professionals alike have migrated *en masse* from analog film to digital formats. Since the first repository specifications for digital photography were drafted in 2006, we've seen digital photography overtake and dominate the field, largely overtaking film as a common medium for the capture of still images.

Of course, new objects will continue to be created using traditional film, and there is no foreseeable end to the creation of objects that originate on paper, film, or other analog recording format, even if those formats are relegated only to niche interest groups. To that end, the repository has established and refined a set of clear and concise standards that serve to acquire and preserve digital facsimiles of analog photographs, books and similar items.

Even so, digital photography brings with it new challenges and different capabilities than our existing core set of scanning digitization standards can support. As a result, an entirely separate set of standards dealing exclusively with digital photography and separate from those that support scanning must be defined and adhered to.

Emerging shifts to digital photography

While we have long heard that film's days are numbered, few have truly believed it until very recently. Digital photography has taken more than 12 years to mature, since the introduction of the first mass produced digital camera (the Apple Quicktake) in 1994. For a majority of this period, the switch from film to digital was largely relegated to early adopters, and broadly shunned by professionals who insisted film was here to stay. Within the last decade however, the quality of the hardware available as well as the introduction of professional grade software tools has not only swayed general opinion of digital photography, but has permitted digital photography to become a driving factor in the fate of most corporations in the field. Additionally, a number of very recent events has permanently and irrevocably spelled out that film's days as a dominant medium are numbered:

- **October 12, 2001**: Polaroid, Inc. files for bankruptcy. This is often seen as the watershed event for the decline of analog formats. Development of instant film formats stops, and while the popular Land Camera and a few other versions of Polaroid film survive, a wide array of other formats were discontinued.
  (Since 2001, Polaroid has been resurrected, filed for bankruptcy yet again, and the instant film formats discontinued. At present, private enthusiasts have attempted to revive Polaroid instant film through independent efforts.)

- **2001 – 2006**: Kodak has progressively discontinued a number of film formats, though it has stated it will aggressively pursue the continued manufacture of conventional 35mm and APS film. Additionally, Kodak announced in 2004 that while it "is, and will remain, committed to manufacturing and marketing the world's highest quality film," it is ending production of film cameras.

- **January 7, 2003**: Konica and Minolta, once both strong names in the film and film camera businesses respectively, announce they will merge to form a single company. This is largely viewed as the result of dwindling revenues from analog format sales, as both companies seek to share their digital technologies to strengthen their position in this market.

- **December 2005**: Kodak announces that for the first time, revenue from digital cameras and digital storage media has exceeded revenue from film-based sales.

- **January 11, 2006:** Nikon announces that is has discontinued all but two 35-mm Single Lens Reflex (SLR) cameras: The F6 and the FM10. It also announced it will discontinue the manufacture of all large format analog lenses, and all but nine interchangeable lenses to support the F6 and FM10. In addition, Nikon's photography division announces it will focus almost exclusively on the development of its digital product lines.

  As of 2010, the Nikon F6 and FM-10 continue to be manufactured, although the FM-10 is made by Cosina, and rebadged as a Nikon.

- **January 19, 2006:** Konica Minolta announces it will exit the photography business altogether, discontinuing both analog and digital film camera lines. It will sell its technology to Sony, which has indicated it will continue to support existing Konica Minolta digital camera lines, and develop new lenses compatible with the K-M lens mount.

- **July 22, 2009**: Kodak announces that it has manufactured its final batch of Kodachrome film after 74 years of production. Kodachrome was well known for its longevity and color stability. The last stocks of Kodachrome film have an expiration date of December, 2010.

- **January 2010**: Canon exits the analog film camera business by quietly discontinuing the manufacture of the EOS 1v. While remaining stocks of new EOS 1v cameras can still be purchased at retail stores, and while most lenses Canon makes for its digital cameras will still work on the film EOS line, all of the cameras Canon currently makes are digital-only.

- As of this year, digital images are estimated to account for 90 percent of all professionally taken photos according to market research firm InfoTrends.

At the same time that film-based companies are seeing the need to adapt or perish in the digital realm, digital cameras have improved dramatically in image quality. While there was once a time where the idea of using digital photographs to preserve images and keep permanent records was laughable, manufacturers are now producing affordable digital cameras – some aimed at entry-level users - that can meet or exceed the image quality produced by some 35mm film types.

These events point to one conclusion: analog film will continue to serve a greatly reduced role in the field of both amateur and professional photography as time progresses. While it is unrealistic to say that film will altogether become extinct, the prevalence of the common traditional formats (35mm, 110) are on the decline. It is very likely that film will be relegated to a limited range of formats for special-purposes applications and niche audiences, while more common general-use and utility-based photography will overwhelmingly shift to digital.

The need for baseline standards

The shift to digital photography has not been easy, and has been fraught with many painful lessons on what constitutes acceptable image quality. Indeed, early digital camera models produced

images that were barely acceptable even for computer equipment of the time, much less for print media. Nonetheless, attempts were made by early adopters to use the technology for permanent preservation, and the results are that the digital images produced are unacceptable for viewing.

Indeed, for our purposes, digital cameras are only now being produced that can match the exacting standards that RUcore has laid out for acceptable, preservation-grade images. As the quality has improved, so has the acceptance and adoption of this hardware for general use photography. This is an important turning point for RUcore, as although our repository has a number of professional grade images in our collections, the majority of the photographs we have preserved thus far are often donated family photographs, amateur stills, and images that were generally produced using consumer equipment. As a result, we can expect that in the not-too-distant future, we may be expected to preserve amateur as well as professional digital images that are deemed to capture images and moments that are preservation-worthy.

In preparation for this, it is essential that RUcore adhere to a standard for which we will accept born digital images for inclusion in the repository.

Why have a separate standard from those for scanning photographs and documents?

At first glance, it might seem very easy to take the established standards for photograph and document digitization, and simply apply them as-is to digital photography. Indeed, the two processes share some similarities, and some of the requirements established for digitization should serve as the basis for establishing comparative standards for born digital still images. However, there are a few key differences between digital photography and analog digitization that make a broad application of a single standard impractical. Consequently, the two workflows need to be viewed from different paradigms to fully understand them and appreciate their differences.

Perspective is everything: digitization terms redefined

The best way to understand the differences between digital photography and digitization workflows is to view their intended purposes.

Digitization, or simply scanning, is intended to take an object recorded on an analog medium such as film, slides or paper. From this, we use an array of equipment and software to create a digital facsimile, with the intent of making the digital form represent the source object as accurately as possible. Consequently, the workflow, specifications and terminology are centered around this process.

Digital Photography on the other hand, is a process where the digital form *is* the primary, original storage medium. With digital photography, there is no physical medium that can accurately be described as the "original." In order for the digital format to take the primary role in recording and preservation, the hardware must be designed differently, and procedures and terminology have to take significantly different characteristics from digitization.

These differences in purpose and perspective result in important variations in how images are acquired and described:

Resolution: PPI vs. Megapixels: The most important difference between digitization and digital photography is the issue of resolution. Those familiar with digitization have grown accustomed to expressing resolution in terms of pixels per inch (ppi). This is because for digitization purposes, resolution is a function that expresses how accurately a scan will replicate the original. the higher the ppi, it is presumed, the higher the quality of the resulting digital image will be.

Digital photography, however, limits the relevance of ppi in terms of creating the original photograph. As image sensor sizes can vary greatly from one camera to the next, it is possible for two different camera models to arbitrarily assign widely different ppi values to their images, yet still produce

digital images that are of comparable overall quality. In such a case, ppi only comes into play when a user wishes to print the digital image, in which case this value can be changed at will to suit the user's needs. As a result, the value of importance in digital photography is not how many pixels per inch make up an image, but the overall **pixel count**, or number of total pixels, that are used to represent the image. With current technology, this value is frequently expressed in Megapixels (MP).

Unaltered Originals: RUcore places the utmost importance on the ability to have an archival digital master, that is unaltered or unedited in any way. This requirement ensures that we can refer to this original at any time, should any edits or calibrations we perform on our derivate presentation versions of an object become unsuitable for display as technology changes. Producing such images are relatively easy when digitizing analog formats. The matter becomes trickier, however, when dealing with digital camera equipment.

Born Digital File Formats: JPG, RAW Image file formats and the unique challenges they present

To be sure, no single digital camera architecture will suit every photography application and so, camera vendors design and construct a vast assortment of digital cameras that vary in size, resolution and capability. A major challenge for dealing with digital photography is the diversity of equipment that is out in the field, and the resulting file formats that they generate.

Entry-Level Consumer Digital Cameras pose the greatest issue because they typically output files using the JPEG file format, with very lossy compression. To their credit, such cameras permit beginners and casual users to capture important and even historic moments with a minimum of effort and skill, and a great deal of archived content would not exist without casual photographers using such equipment, where more advanced and skilled photographers are simply not present. However, their ease also presents a disadvantage: entry-level cameras heavily process the images the capture, and the resulting image files are suboptimal for archival purposes without, at the very least, a file format change to an uncompressed TIFF format.

"Pro-sumer" and Professional Cameras typically provide the option to process and compress captured images into JPEG files similar to the consumer counterparts, but also tend to provide an option to yield *camera raw image files*. A camera raw image file contains minimally processed data as retrieved directly from the image sensor of the digital camera. Raw files are so named because they are not yet processed and therefore are not ready to be printed or edited with a bitmap graphics editor. Normally, the image is processed by conversion, where precise adjustments can be made before creating a "positive" file format such as an uncompressed TIFF or JPG file. Similar to a film negative, a raw digital image may have a wider dynamic range or contain more color information than can be provided using currently used file formats for presentation and access (TIFF, JPG, etc.), and preserves most of the information of the captured image. The purpose for a raw file is to achieve minimal loss of image data obtained from the sensor, and the conditions surrounding the capturing of the image (the technical metadata). In the field of photography, there is a pervasive, erroneous belief that RAW represents a single file format. In fact there are hundreds of raw image formats in use by different models of digital equipment, and the formats can vary from one vendor to the next, and even among different camera models made by the same manufacturer.

To get around the issue of non-standard and widely-disparate raw image formats, a standardized open file format, developed by Adobe Systems, Inc. and called "Digital Negative" (DNG) was developed in 2004, and is updated regularly with backward comaptibility. DNG is based upon the TIFF image standard, but encapsulates the additional sensor data in most proprietary raw image formats. In addition to Adobe software, the DNG file format is accessible and can be read by over 40 additional 3rd-

party software packages across Windows, Mac and linux platforms. Because of this, RUcore tends to prefer capturing and preserving raw image files that have been converted to DNG, as these represented minimally-processed image files in an open, well-documented format that preserves not only an uncompressed digital image, but a wealth of associated technical metadata.

Recommended Born Digital Imaging Standards

Taking into account the aforementioned considerations, RUcore strives to adhere to the following recommendations for born digital still image content:

Resolution Requirements:

- **For entry-level consumer cameras: *Minimum* of 7.0 effective Megapixels (MP),**
  ***or* 5.0 Megapixels if the camera has a "High Dynamic Range" (HDR) capability built-in.**
  - Most entry-level "point and shoot" cameras heavily process and compress photos taken with them, introducing artifacts. Additionally, smaller imaging sensors in these cameras contribute to sensor noise. The high minimum resolution is necessary to help overcome these issues.
- **For "Pro-Sumer," bridge cameras, and professional dSLR cameras:**
  ***Minimum* of 6.0 effective Megapixels (MP)**
  ***or* 5.0 Megapixels if the camera has a "High Dynamic Range" (HDR) capability built-in.**
  - The resolution requirement for non-entry level cameras is lower because it is possible to obtain unprocessed, uncompressed images from these cameras, generally yielding better results even with less image information.
- **Additional considerations for both classes of cameras:**
  - Use of "total" or "interpolated" pixel counts to meet the standard are *not* acceptable, when the effective count is below the minimum.
  - A camera will *not* qualify as preservation-grade if it uses interpolation to reach its advertised resolution.
    - Example: A manufacturer advertises an extremely inexpensive digital camera capable of producing 10MP images, however the fine print indicates the camera is only equipped with a 3MP sensor. This camera is in fact interpolating a 3MP image to 10MP, and is not acceptable for preservation purposes.
- ***Minimum* 8 bits per channel (24-bit color)**
  - The camera should be capable of producing images using the sRGB palette.
- **The equipment *must* be capable of producing images with pixel dimensions of at least 3,000 pixels on one side.**
  - Example dimensions: 3504 x 2336; 3072 x 2902; 3872 x 2592; and 3264 x 2448 are all acceptable.
- **The equipment must be EXIF compliant, version 2.0 or later.**
  - EXIF compliance ensures the camera will embed metadata into the image file that details program modes, exposure settings, lens type, and other relevant information.

Image Format Requirements:

- **For consumer digital cameras: A direct copy of the JPG output file, without any post-processing.**
  - When possible, this JPG image will be directly converted to a TIFF file, without *any* changes to resolution, image quality, brightness/contrast, levels or other aspects.
  - An edited copy of a digital image is permitted if the edits are the direct result of the photographer's intent to present the image with such modifications for artistic effect. When permissible, an unedited "master" should also be preserved, but will not be made publicly accessible or viewable.

- **For ProSumer and professional cameras: The equipment should be able to produce images in RAW format.**
  - RAW image format ensures that the images produced by the camera are unprocessed, unedited and uncorrected.
  - The camera should either be able to produce image files conforming to the **Digital Negative (DNG)** file format, *or* interface with software that can export a DNG file from the camera's proprietary RAW format.
    Common software packages for this purpose: Adobe Photoshop, Adobe Lightroom.
    Additional listings of 3rd-party software packages can be found at http://www.adobe.com/products/dng/supporters.html
  - In addition to the DNG, a derivative TIFF file will be created and stored as a preservation format, through which presentation JPG, PDF and Djvu or Jpeg2000 images will be created for access by the public.
  - DNG permits the photographer to specify image and lighting adjustments, while not destructively altering the original image.

- **Alternately, the equipment should be able to produced uncompressed TIF images.**
  - Uncompressed TIFs can be used as an archival master, but bear in mind that DNG is the preferred format. Care should be taken when using TIFs to ensure that no image processing occurs to the TIF file, beyond what the camera performs internally. The same considerations will be made for artistic adjustments as in the treatment of camera-produced JPG files.

Other Considerations:

- **Image quality:** the equipment must be able to produce images with a minimum of sensor noise, and with optimal and accurate color reproduction. Such criteria is subjective, but generally most common photography equipment from major vendors will yield acceptable images as long as they meet the above specifications.
  When possible, a non-exclusive list of tested and known-good cameras will be maintained and made available.
- **Image stabilization:** If you choose a camera or lenses with Image Stabilization (IS), be certain the IS engine is of an "optical" variety, not "electronic" or "virtual." Optical IS uses floating internal lens optics and gyroscopes to ensure a steady image if the camera is moving. Electronic/Virtual IS uses software-based image editing and interpolation to artificially render a steady image.

- **Images taken from cameras not meeting the preservation spec:** It is inevitable that events will occur where images we wish to preserve in RUcore will be captured by cameras not meeting the above specifications. In the absence of better quality images, such images can be accepted by RUcore on a case-by-case basis, in which the RUL Digital Data Curator or the Digital Preservation Task Force will need to evaluate the images and determine the best course of action. It should be stressed however, that the viability of such images cannot be guaranteed and any preservation efforts will be done on a "best effort" basis.

RUcore
Rutgers Community Repository

Sound Objects:
Recommended minimum requirements
for preservation sampling of audio

Introduction

This document will set forth two standard requirements for audio. One will establish a minimum and recommended sampling rate – the quality level at which the audio is digitized – for the digital audio masters and presentation copies. The second standard will recommend specific file formats for the preservation master and derivatives, for implementation into the Workflow Management System (WMS).

Although the standards will be different, the philosophy behind preservation and presentation will be same as for all other object types. It will be mandatory to archive an uncompressed archival master, to ensure an object of the highest quality is preserved. Additionally, a small but diverse number of presentation copies will be archived as well. These presentation copies are to be stored and accessible in formats that the end user will find easy to play back, and will be "low-bandwidth friendly" whenever possible, allowing users with slower internet connections to have access to these objects as well.

Sampling and Digitization Rationale

As with all other objects, obtaining a high quality sample of the original for preservation in RU-CORE will assure the best chance of long term preservation without having to go back to the original source for a resample in the future. This will also allow us to ensure that the presentation copies provide a comparatively high fidelity that sacrifices little in quality. In the digital realm, audio is represented by a digital sampling at a set frequency, to obtain a granular but reasonably accurate representation of the analog original. Sampling is the process of converting a signal (e.g., a function of continuous time or space) into a numeric sequence (a function of discrete time or space). The higher the sampling rate – it is assumed – the more accurate the digital representation will be.

For audio, there has been a wide practice of following the *Nyquist-Shannon Sampling Theorem*, a doctrine which is used to assert that 44.1kHz is an acceptable minimum sampling rate for all audio. This belief is based on the established fact that most human ears perceive sound up to an upper frequency threshold of 20,000Hz, and sampling must occur at twice the upper limit to achieve an acceptable digital copy. Consequently, a number of digital recordings, including CDs, adhere to this standard sampling rate (thus the term "CD Quality" is attributed to this sampling rate).

This 44.1kHz sampling rate is not without its detractors. Over time, audiophiles have consistently complained that they perceive a loss of fidelity when analog recordings are digital remastered to CD Audio. While some audio experts have insisted that these complaints are based on purely psychological factors, there is some support for a need for a higher sampling rate. There are inherent risks in losing quality to the sampling process, causing a degradation that is not accounted for in Nyquist. However, a higher sampling rate may be able to compensate for these sampling losses.

As a result, the standard set forth accounts for the CD-Audio minimum sampling rate and accepts it as a minimum, while recommending a higher level whenever the opportunity to sample at a better rate presents itself.

Recommended Standards for NJDH and RU-CORE Audio Sampling

- **Minimum sampling rate: 44.1kHz 16-bit (CD Audio)**
  This is the minimum acceptable rate to ensure a good preservation master. Most Compact Discs (CDs) are mastered at this rate. As such, all audio obtained from CDs will be archived at this rate. Additionally, 44.1kHz is a suitable sampling rate for RU-Core partners when mastering recordings of spoken-word speech (i.e. interviews, speeches, press conferences and lectures), that are not accompanied by high-fidelity sound or music.

- **Recommended Sampling rate: 96kHz, 24-bit audio**
  This is widely considered an ideal rate for high quality audio recordings, including DVD-Audio. For most audio formats, this sampling rate is the maximum sampling rate that also supports Quad (Dolby 4.0) and Surround (5.1) audio. When repository content partners are making a first generation sample of musical or high-fidelity recordings from an analog master, it is recommended that this sampling rate be used whenever technically possible.

- **High Level (Maximum) Sampling rate: 192kHz, 24-bit audio**
  This sampling rate is often touted by audiophiles as one of the best sampling rates to work with in the editing of audio recordings and creating master samples. However, this format is generally not supported in current mass-produced formats for Quad or Surround sound. As such, recordings sampled at this rate should be limited to Mono or Stereo recordings. In general, this sampling rate, and higher rates, are recommended if there is a reasonable justification for using such a high sampling rate, and it is believed that the 96kHz rate will not be sufficient for accurate reproduction of the original sound.

Recommended File formats for preservation and presentation of audio objects

The following formats are recommended for the preservation and presentation of audio.

- **For Preservation: Standard WAV or Broadcast WAV Format (BWF)**
  BWF is an extension of the popular WAV audio format. It was first specified by the European Broadcasting Union in 1997, and updated in 2001. WAV records audio using Pulse Code Modulation (PCM), the industy standard method for digitizing audio and is used in CDs and DVDs.
  The stated purpose of these two file formats is the seamless exchange of digitized audio between different computer platforms. BWF also specifies additional metadata, allowing audio processing elements to identify themselves, document their activities, and permit synchronization with other recordings. This metadata is stored as an extension chunk in an otherwise standard digital audio WAV file.

- **No compression of archival master is recommended**
  As of this writing, the Audio and Video Standards Working Group recommends that no compression of the preservation master occur. While there are some lossless compression formats available (e.g. Shorten and FLAC), the open source formats that are currently available are not mature, nor do they have a large enough user base to justify their use. Doing so may expose the repository to the risk of being unable to later decompress and access these masters if at some point in the future, support and development for the chosen compression scheme is abandoned. However, the working group does recommend that the issue of lossless compression for archival masters be re-assessed at a later date, to determine whether an open standard is more widely accepted, likely to be readily available and supported for the foreseeable future, and suits our needs.

- **For presentation Audio: MP3 or Ogg Vorbis, using Variable Bitrate (VBR) encoding**
  Both file formats are widely used by computer end users and supported by most popular audio playback hardware and software.

  MP3 enjoys wider acceptance, but is a format that is encumbered by proprietary compression algorithms. However, current licensing restrictions indicate that we would not be required to pay royalties for non-commerical, non-profit-generating use. Ogg Vorbis, while not quite as widely accepted, still enjoys support from the audiophile community and is an open source format, without any proprietary encumberances. The drawback however, is that Ogg Vorbis is not natively supported by common players such as Windows Media Player, Apple Quicktime, and some mobile devices.

  For this reason, MP3 is the current standard presentation audio format for RUcore.

Evaluating collection objects that do not meet standards

The working group recognizes that there has been a period of at least two decades where digital audio has been recorded and exists prior to the establishment of these guidelines. It is important to acknowledge that there is a prevalence of digital audio objects that may be of immense value to repository partners, but for which there is no analog master available and the best digital master may not meet our established digitization standards.

In light of this, it is important to stress that the standards we have established are recommendations, and must not be the only criteria for accepting or dismissing a potential audio object. While we believe it is of the utmost importance that collection partners strive to meet the standards in order to ensure longevity of their collections, the advisory committee should consider the overall content and value of the collection before making a decision as to its inclusion. In particular, the committee may want to evaluate:

- The playback quality of the objects, and whether the audio quality can subjectively be deemed acceptable in spite of not meeting standards.
- The importance, prominence, and significance of the content
- Whether further degradation of the content can be inhibited by storing the object as an archival master, or converting an object with lossy compression into a lossless format.

If the advisory committee decides that the benefits of storing an object or collection into the repository outweigh its lack of standards compliance, then the standards can be waived for that object or collection. However, in doing so, the point should be stressed to the collection partner that long term preservation of the object *cannot* be guaranteed. While the repository and the team supporting it will put forth its best efforts to sustain the collection, the collection partner should be made aware that the chances of losing the object to format obsolescence or degradation of integrity are greatly increased because the object has not been digitized to our specifications.

RUcore
Rutgers Community Repository

Video and Moving Image Objects:
Recommended Minimum Standards
For Archival and Presentation Datastreams

Introduction

This document will set forth a standards recommendation for moving images and digital video. In particular, this video object standard will recommend specific file formats for the preservation master and derivatives, for implementation into the Rutgers Community Repository (RUcore) and projects using similar architectures, as well as recommend sampling rates and specifications for presentation derivatives.

As with all other standard types established thus far, it will be mandatory to store and preserve an archival master, to ensure an object of the highest available quality is maintained for digital preservation. Additionally, one or more downsampled and compressed presentations copies will be made available for end users wishing to access these objects online. These presentation copies are to be stored and accessible in formats that users will find easy to play back, and will use file formats and codecs that are compatible with multiple computer platforms, using established industry standards.

Sampling and Digitization Rationale

The handling and preservation of digitized moving images presents a unique challenge to digital repositories. Presently, uncompressed digital video demands an extremely large amount of storage space, and produces incredibly large files. Yet, the need to store an uncompressed or reliable lossless-compressed object is paramount to ensure its longevity. While it is recognized that work continues in perfecting lossless video compression standards, we feel that these codecs are not mature enough and have not yet reached a critical mass in terms of user base and supporting software to implement in place of an uncompressed stream. We remain open to revisiting this stance in the future.

We also recognize with the growing convergence of digital devices, and the prevalence of smaller video capture equipment, there will be an increasing amount of digital content which is born in a compressed digital format. Such cases will pose long-term preservation challenges depending on the file times, video codecs, resolution and compression levels used. When such video is slated for inclusion into RUcore, a case-by-case condition analysis will occur; best efforts will be made to store the native format as an archival datastream; and when necessary, a converted copy into a designated stable format will also be stored with the archival datastream.

In spite of the present need to store an uncompressed stream when digitizing from an analog master, it is obvious that delivering such an object to end users would be impractical given current average connection speeds. Consequently, there is an additional need for downsampled, compressed presentation formats for video objects, more than any other object type addressed by the repository.

As always, the guidelines presented here are recommendations, and there may be cases where judgment calls will need to be made about objects that would be better preserved by modifying the recommended guidelines for this purpose. In particular, the digitization team has not yet digitized film archives, and as such those formats will need to be analyzed for the best possible digitization settings. The Digital Data Curator, as well as the Digital Preservation Task Force, should be consulted for guidance when such adaptations are required.

RUcore Media Standards Working Group:
I. Beard, I. Bogus, E. Gorder, N. Gonzaga, B. Nahory, R. Sandler

Recommended Standards for NJDH and RUcore Video Digitization

For analog preservation masters (when possible):
File format: *Uncompressed, Full Frame Video (AVI file format) or DV Source for digital video.*

Frame rate for analog Standard Definition (SD) video, NTSC: *29.97 frames per second, 640 x 480 resolution (assuming square pixels). 4:2:2 quantization, 25MiB/s data rate.*
We recognize this sampling scheme as the best practical standard to ensure a good preservation master of analog SD video archives, and will be the most common digitization sampling rate for objects that come to us as SD analog video. This standard is based on our experiences with digitizing videotaped objects.

For Digital objects (i.e. DV/HDV), including high definition video: *Use and preserve same frame rate, resolution and bit rate as the original.*
For born-digital video objects such as DV or MPEG-2, the logical course of action is to preserve the exact specifications of the original. It will not be wise to downsample the original as that will cause a loss of object data, and no improvement in quality will be gained from upsampling.

All other objects: Make best effort to preserve frame rate and resolution of the original content. The goal in digitizing the various analog formats that may come to us will be to create a digital master file that preserves the content of the analog original as accurately as the digital media permits. A wide degree of flexibility and some experimentation may be required to determine accurate settings for each unique case.

Presentation video files:

- **One streaming/progressive downloadable video clip:**
  - **MPEG-4 H.264 video (.MOV, .M4V, .MP4), encoded for hinted streaming**
  - For 4:3 – Minimum of **640 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - For 16:9 - Minimum of **854 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - Recommended Data rate of **640 kbps minimum, and up to 860 kbps.**
    Use higher bitrates for videos with more detail and greater motion.
  - **Key frames inserted every 30 frames at minimum, or auto-select. This rate should be adjusted when necessary for best results.**

  This recommendation is aimed at balancing the file size, and the amount of bandwidth required to play the video, while trying not to sacrifice video quality. This specification necessitates the use of a broadband internet connection, but is configured so that basic Home DSL or casual WiFi users should still be able to view the content.

  MPEG-4 Video, particularly MP4, is cross-platform and can be accessed by desktop computer users of varying operating systems (Windows, Mac, Linux), using free software and established web standards. H.264 video is also viewable on a multitude of internet-connected mobile devices.

  Starting in late 2010, the MP4 container format is recommended, as this format permits us to use a single H.264 video file to provide service for mobile devices as well as progressive download and streamed video.

Progressive download standard for older objects

Prior to September 2010, the standard for progressive-download presentations videos were as follows, but has since been deprecated with the use of the single-source MP4 spec listed above:

- **If permissions permit: one progressive-download video clip**
  - **Flash Video Format (.FLV), using ON2VP6 Codec**
  - For 4:3 – Minimum of **640 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - For 16:9 - Minimum of **854 x 480 resolution (square pixels), 30 frames per second, multi-pass encoding**
  - Data rate of **512 kbps**
  - **Key frames inserted every 30 frames. This rate should be adjusted when necessary.**

  Our experimentation has shown these output settings to be an ideal compromise, producing a clip viewable at acceptable quality on a computer screen while providing a reasonably manageable file size. Users choosing to view this format will need to download the latest version of a free Macromedia Flash Plug-in, provided by Adobe Systems, Inc.