DIGITIZATION TECHNICAL LEAFLET

Midwest Art Conservation Center
June 17, 2008

CONTENTS

INTRODUCTION

A digitization project can be as simple or complex as an institution wishes, but unfortunately it is never as easy as it might appear. However, digitization presents tremendous opportunities for increased access and added value to intellectual property. To maximize the investments required to make it happen and the available opportunities, a number of issues should be thoroughly considered. Elements of a comprehensive digitization program includes selection criteria, digital asset management, rights management (including intellectual property rights and risk management), standards for metadata and reformatting, and an ongoing digital preservation plan. This document briefly covers the various factors that should be taken into account as an institution is considering a digitization effort.

- A key first step is to determine the potential use and audience of the digitized collection.
- Achievable goals and priorities will reflect the mission of the institution.
- Documentation is vital in creating a sustainable program.
- Rapidly evolving technology make a sustainable digital preservation plan essential.
- Repositories pool resources and alleviate infrastructure investment.

- Robust metadata (any type of descriptive information) greatly increases the value of the digital object and maintains the intellectual integrity.
- The use of uncompressed, non-proprietary file types will best ensure long-term viability of reformatted materials.
- Consultants alleviate the need for the institution to become experts in all media and equipment types and to invest in expensive or obsolete equipment.
- Digitization projects present numerous opportunities for collaboration.
- Talk to colleagues! Most professionals are more than happy to share their experiences to benefit the community!
- Look online! There is an *enormous* amount of information available online. Most institutions have made the selection criteria and best practices available online.

MOTIVATION AND OPPORTUNITY

Motives for digitization include the desire to protect, represent or transcend original objects, alternately, to provide access, preservation, innovation or outreach. Digitization provides tremendous opportunities to preserve perpetually aging collections, reduce the costs of replacing highly used materials (sustaining increased wear-and-tear) or accessing materials that are not readily available or difficult to present. Digital collections can be utilized to promote collections to a wider audience or finding aids may be placed online providing access to non-digital collections. Opportunities for new and possibly un-foreseen uses and connectivity are dramatically increased by the addition of rich metadata to online collections. Additionally, digitization projects often present exciting opportunities for collaboration and resource sharing.

Ideally, an institution will digitize collections when the value added by digitization outweighs the costs of the process. Largely due to continually evolving technologies, an institution must not look on digitization as a finite project, but as a course of action that must be sustainable. *A sustainable strategy integrates the digitization project into the institution's services; focuses primarily on mission-related objectives; is funded through continuing and predictable allocations; and includes a plan for the long-term maintenance and preservation of the digitized objects (Smith, 2001).*

Traditional Materials vs. Digital Objects

(Modified from Frey and Reilly, 2006)

Traditional Materials	**Digital Objects**
Traditional materials gradually degrade over time regardless of the quality of composition or storage facilities.	Digital objects do not degrade – the files are either accessible (read correctly) or not accessible.
Traditional duplication results in deterioration of the quality of the copy.	Digital duplication is possible with zero loss.
To save traditional materials: store in optimal conditions and *never touch again.*	To save digital objects: equipment requires regular maintenance and files must be copied to a new media or format before it becomes unreadable. Evolving technology and planned

	obsolescence make the lifespan of both hardware and software typically shorter than the life of the recording media.
Degradation can never be completely halted, only slowed.	A digital object is frozen - it exhibits no degradation over time.
With the exception of audiovisual media, most traditional materials do not require additional equipment for use.	Digital media ALWAYS requires compatible hardware and software for use, *making a sustainable long-term plan essential.*

PROJECT ELEMENTS

Planning and Selection Criteria

A thoroughly planned and well-documented selection process is essential to an achievable project. It lays the groundwork for processes and standards and will guide decision-making if complications arise. Clear documentation will assist peers and successors and provide justification for sustained funding. Especially when an institution is considering its first digitization project or establishing a digitization initiative, thorough planning is crucial and all elements of a project or initiative should be addressed. Often unrealistic expectations stall the project, creating unachievable commitments for staffing, equipment, infrastructure and funding.

Selection criteria are typically formed by some combination of the following goals: preservation, access, outreach or innovation. Prioritization motives may include:

- Digitize to increase or improve access to the scholarly and cultural record (and create new forms and functionality).
- Create digital surrogates to minimize risk issues and improve intellectual control.
- Create digital surrogates to limit physical contact with fragile, deteriorating or extremely valuable materials.

It is recommended that these three questions be answered and justified before embarking on a digitization project (Hazen, Horrell, and Merril-Oldham, 1998):

1. Does the material have sufficient intrinsic value to ensure interest in a digital product?
2. Will digitization sufficiently enhance access or increase use by an identifiable constituency?
3. What goals might be met by digitizing?

The following steps should be taken over the course of the planning process (CDP Digital Audio Working Group, 2006):

1. Clearly articulate the goals of the project, *in terms of the institutional mission*.
2. Identify the audience for the digital products (is there a demand?).
3. Have a thorough understanding of methods and processes for completing the project.
4. Develop a realistic work plan with phasing options.
5. Identify staffing or outsourcing needs.
6. Create budgets and funding sources to support the project through a series of stages.

Ultimately, a digitization project should be fully incorporated into the institution's mission and collection development strategies. A sustainable effort demands an institutional commitment. To provide long-term benefits, the project may be based on the strengths of the existing collections, augmenting existing digital collections, or filling curriculum or research needs. New and expanded projects are continuously being developed; it is helpful to determine if the project is unique or would augment an existing collection.

>>>Please also see RESOURCES for links to decision matrices, selection criteria, best practices, workflow and other guidelines.

Considerations

Collection Materials

- What is the purpose and benefit to digitizing materials; is there a demand for the digital format? Will the digital format replace or supplement the original; does it supplement an existing collection or fill a need?
- Are they already cataloged; is existing descriptive information (metadata) available? If not, time will need to be spent completing cataloging or creating descriptive records. What metadata or finding aid schema will be used (Dublin Core, MARC, EAD, etc)?
- Does the institution have appropriate rights to provide greater access? If not, may access be granted? Limitations may arise from copyright, donor or other legal issues.
- Are the materials in suitable condition to be handled? Is conservation treatment necessary in order to prepare the objects for digitization?
- Reformatting standards should be sufficient to avoid the re-scanning and re-handling the objects in the future. The quality of the master file (preservation copy) should ideally be as high as achievable based on the quality of the original and as affordable by the institution. There is no advantage to scan at a higher resolution than what is needed to obtain the detail of the original; a high-resolution digitization of a poor quality original increases file size without providing any additional information.
- Derivatives may be of variable quality based on use (print vs. screen-legible, etc.). Scholarly access may demand high-resolution products while reference copies may use lesser resolutions. Symphony recordings may demand higher quality audio formatting than spoken word.
- Metadata (any form of descriptive information) provides value-added information, increasing use capabilities and opportunities. Providing transcripts or captioning for audiovisual materials or search capabilities for text-based materials adds immense value, but also adds significantly to the schedule and cost.

Staffing

- Does the institution have experienced staff? Will current staff be trained or will the institution hire new? Use permanent or short-term staffing? Is the digitization initiative sufficient to employ, at a minimum, a project manager, especially if there will be an expected turn-over of short-term staff?
- Will a vendor be used? Will the work be performed on-site or off site?

- What forms of quality control will be established, internally and for vendor relations? Comprehensive documentation is crucial to a consistent ongoing project.

Equipment and Infrastructure

- Does the institution own the appropriate equipment? Is the institution able to provide continued support for the equipment? In addition to purchase costs, annual costs of maintenance and upkeep are often forgotten. Potential investments include workstation, digital storage and backup system, scanners, digital cameras, audiovisual play-back devices, analog-to-digital converters, processing software, etc.
- Will the digital objects be made available online or internally at a workstation/reading room? What modes of access features will best help the users? What search features and user interface? Online availability requires appropriate server components, bandwidth and infrastructure (which may be alleviated by using an off-site repository). Internal access reduces many opportunities, but may also be used as an initial step in a phased plan. If limited access is necessary, at a minimum, finding aids should be made available electronically.
- Swiftly evolving technology requires that collections be maintained to accommodate new hardware and software systems. File formats may need to be converted to meet future standards; digital storage and delivery media will likely need upgrades/replacement.

Rights Management

A primary goal of digitization projects is to increase access to the intellectual information of objects. As a result, questions regarding intellectual property rights are integral to the digitization process. Common questions include: is the work protected, does the institution have permission to make the work digitally available, is the use a "fair use", what rules apply, and what happens if permission is not granted? Institution-produced and gifts may present unforeseen issues. For example, permission to interview doesn't mean permission to digitize or make available online; permission must be obtained from both the interviewer and interviewee. A decision tree approach is recommended to reduce legal risk and simplify the burden of obtaining permission.

The New Jersey Digital Highway has identified six key steps to create a practical and defensible rights management strategy:

1. Understand copyright.
2. Identify resources that are clearly in the public domain.
3. Obtain permission for resources that have an identifiable copyright holder.
4. Develop a "risk management strategy" based on a continuum of resources that range from clearly copyright protected to clearly in the public domain.
5. Clearly identify appropriate use of digitized resources for your users through an open license for your users.
6. Develop and publish a copyright policy for your organization.

>>>Please also see RESOURCES for links to copyright tutorials, flowcharts and guides.

Digital Asset Management

Digital Asset Management (DAM) is the process allowing institutions to manage and maximize digital resources. A comprehensive DAM system consists of a **repository** to contain digital

resources and metadata, an **infrastructure** to manage and preserve the digital resources, and **search capabilities** to enable users to discover and access the digital resources.

DAM policies provides lifecycle management of the digital object and can be as simple as creating a file directory containing the digital object and accompanying metadata. The system may consist of a custom-built database or a provider may be used; the system may be hosted and managed in-house, by a provider, or a combination or both.

Common providers include:
ContentDM http://www.contentdm.com/
Corbis Media Management http://pro.corbis.com/
Gallery Systems http://www.gallerysystems.com/
Luna Imaging http://www.lunaimaging.com/
Fedora Commons http://www.fedora-commons.org/
Drupal http://drupal.org/
ImageFolio http://imagefolio.com/
Interwoven http://www.interwoven.com/
Minisis http://www.minisisinc.com/

>>>Please also see RESOURCES for links to best practices and asset management.

Repositories

Repositories are shared services or platforms that collect, manage and distribute digital objects, and can be thought of as a digital archiving service. They may be created and used by a single entity, such as Macalester College's Digital Commons (http://digitalcommons.macalester.edu/), regional, such as the Minnesota Digital Library (http://www.mndigital.org/), or "community" oriented, such as DSpace, which represents scholarly works (www.dspace.org). Not only should repositories be leveraged for the long-term preservation and distribution of digital objects, they often provide resources for best practices.

>>>Please also see RESOURCES for links to repository attributes, responsibilities and best practices.

Digital Preservation

Digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change. Digital preservation applies to both *born digital and reformatted content.* Digital preservation policies document an organization's commitment to preserve digital content for future use; specify file formats to be preserved and the level of preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information (PARS, 2008).

There are three primary concepts associated with digital preservation: refreshing, migration and emulation. **Refreshing** involves moving a digital file from one physical storage medium to another, avoiding decay or obsolescence of the storage medium (making the data irretrievable). This is what happens when files are moved during computer or server upgrades. It also prevents loss of data on media like 8-inch floppy disks, which no longer have components on new computers. Refreshing does not address any issues related to evolving file formats. **Migration** converts a file to an updated file format, such as moving a WordPerfect file to Word 3.0/5.0 to

Word 97, etc. However, there may be loss of a certain amount of information, such as text formatting. Conceptually, there will be a limited number of formats in the future. **Emulation** works to develop software applications that would mimic every obsolete and current application (being able to open and use any file format) and run on any computer environment. Periodically refreshing storage media and using non-proprietary and uncompressed file formats (whenever possible) are the best plan for the continuing availability of digital objects.

>>>Please also see RESOURCES for links to best practices.

Approach & Process

There are four common approaches to digitization projects: best practices, scientific, artistic and user. There is not a single correct approach and projects will typically include elements of all four. **Best practices** create standards that best fit the project based on media, scale, institution and goals. The **scientific approach** aims for benchmarks – objective, calculated measurements of image and text quality. The **artistic approach** will either render a faithful representation of the original's appearance or render the intent by removing blemishes and accidents. The **user approach** creates standards to meet preferences and demand, whether on-line or off-line access, high resolution, print-on-demand, etc. Best practices and standards serve as the most demanding and critical users.

Whatever the approach, digitize once - use many times. The goal is to avoid the re-scanning and re-handling of the original materials. The master file should be the highest quality that the original will allow (and that the institution can afford) and capture all "important" information to meet foreseeable needs. For example, on a map, the smallest important piece of information might be a very thin line or tiny text that is difficult to read. That said, there is no benefit to creating a high-resolution scan if the original is of poor quality; this only creates a large file with no additional valuable information.

When at all possible, use uncompressed, non-proprietary file formats to help ensure long-term viability of the reformatted digital objects. Ideally, file formats will not be bound to a particular carrier (whether software or hardware/storage device) and will not be constrained, by passwords or other usage restrictions, to prevent backup methods and disaster recovery operations.

Following are the Library of Congress guidelines outlining the minimum preservation requirements for the Preservation Reformatting Division. They are intended to be baseline standards that can be expanded at any time.

- High-quality master digital files that allow for a broad range of future use,
- Completeness of original work and its digital reproduction,
- Minimal levels of bibliographic description according to LC policies,
- Searchable full-text (at least "dirty" OCR) with minimal encoding (when appropriate), or other modest access aids such as guides, and
- The use of relevant standards and best practice, whenever possible.

Quality control

In addition to well-documented selection criteria, a formal quality control program is essential to a consistent and successful project. It may combine automated activities such as the verification

of resolution and file size with manual checking of color or audio fidelity. Quality should be sufficient to meet the project goals (and, ideally, future needs) and that the original materials only be handled once.

METADATA AND MEDIA STANDARDS

Metadata

Metadata is "data about data." All objects have features regarding content (intrinsic), context (extrinsic) and structure (associations), which can be reflected in metadata. In digitization projects, metadata is often divided into four common categories: descriptive, administrative, technical, and structural.

- Descriptive: describing the intellectual content
- Administrative: information regarding ownership and rights management
- Technical: features of the digital object
- Structural: describing the relationship between multiple files (as in a series or set)

Metadata provides value-added information to describe, arrange, track and otherwise enhance digital objects. It provides intellectual and physical access to content and may include indices, abstracts, finding aids, acquisition and accession records, and catalog records. It can be created from existing records, by automated computer processes, or by newly gathered information.

Controlled vocabularies and encoding systems are essential in allowing digital objects to be found and range in levels of detail and structure. Examples are Library of Congress Subject Headings (LCSH), Art & Architecture Thesaurus (AAT), Categories for the Description of Works of Art, online public access catalogs (OPACs), MAchine-Readable Cataloging format (MARC), Encoded Archival Description (EAD) and Dublin Core Metadata Element Set (DC). Standards ensure interoperability among objects, institutions, and communities.

Careful and comprehensive metadata will both return the highest "investment" and allow the greatest level of information management in the short and long term. Robust metadata provides an infinite number of ways to meet the needs of both traditional and non-traditional users. Metadata allows the maintenance of intellectual integrity, which is considered the primary characteristic of digital objects.

>>>Please also see RESOURCES for links to metadata descriptions, guidelines and best practices.

Search capabilities for text-based objects

Text-based documents may be digitized to facilitate re-use, distribution, or textual and linguistic research or analysis. The institution will need to determine if the objects are to be utilized for the textual content or if the image must be captured to faithfully convey the original object, which often contains contextual information. *The same concepts can be applied to audiovisual materials.*

Depending on material, users, cost and project goals, the institution may utilize "simple" page images, text and image, or fully encoded text. Using **page image only**, the original object is

viewable, but text is not searchable. Minimal metadata is necessary for online accessibility and this is often most appropriate for an online exhibit. When using **text and images**, "dirty" (uncorrected) Optical Character Recognition (OCR) may be used, making it possible to attach searchable text to the image, facilitating greater access and usability. **Fully encoded text** is the most accurate, flexible and functional. Unfortunately, the costs of creating fully encoded files may be up to four times higher than scanning texts as images and manually providing limited metadata.

Textual information may be obtained by using OCR software or manual transcription. **Optical Character Recognition (OCR)** is most easily utilized when the source material is clean, crisp and consistently formatted, uses a modern typeface, and can be fed through a sheet-feeder. Accuracy rates are usually much lower than transcription and the text will still require editing and encoding. **Transcription (typing or re-keying)** should be considered if the source material is rare, fragile, oversized (difficult to scan), includes images or special characters, or handwritten. Advantages are high accuracy rates and the ability to add basic encoding at this stage. However, both time and costs may be high and clear project management and quality control procedures must be in place.

In order to make the text accessible online, the markup (**encoding**) of the text file is another important consideration. Simple HTML provides little more than website display and falls short of sustainable practices. Best utilizing the effort, the text will be encoded in a content-based markup schema. The Text Encoding Initiative (TEI) is an international and interdisciplinary standard developed to represent many types of text for online research and teaching.

Text-based and Image Media

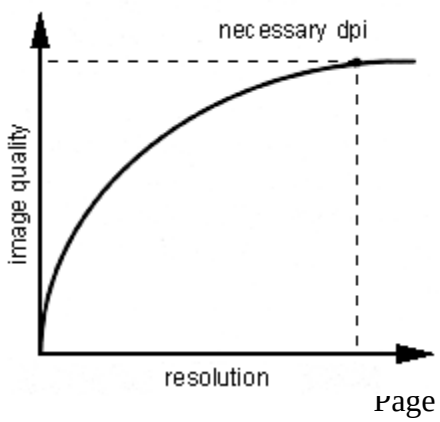
Text and image media include reflective and transparent media such as photographs, negatives, books, manuscripts, paintings, drawings, etc. Digital imaging is often considered the least complicated of digitization projects. Initiated over thirty years ago, there is a wealth of published information and established standards. Minimum components to text and image reformatting are a scanner or digital camera, a processing computer with image editing software, digital storage media and, if applicable, software to complete OCR and encoding.

File formats

TIFF (.tif) is the "de facto" standard for uncompressed files, nearly universal in digital imaging best practices. **JPEG** (.jpg) files use lossy compression. It is not recommended for master files, but may be utilized for certain types of derivatives, such as thumbnail or reference images.

Quality standards

It is important to recognize that higher resolution is not always better – there is a point where the quality curve flattens and higher resolution does not improve image quality. The key is to scan at an appropriate resolution to achieve the project goals / purpose, capturing the smallest piece of valuable information. Ideally, the quality would be sufficient to only require one handling of the original.



This chart is taken from Cornel University's online digital imaging tutorial.

Typical guidelines for text-based materials

- Master files:
  - Bit depth: 1-bit bitonal, 8-bit grayscale, to 24-bit color.
  - Spatial resolution: 300–600 ppi, using 400 or higher if intending to perform OCR.
  - Spatial dimensions: usually 100% of the original.
- Access copies may be constrained to 150 ppi and 600 pixels across the long dimension.
- Thumbnails may be constrained to 72 or 96 ppi and 150–200 pixels along the long dimension.

Typical guidelines for photographs, maps and other graphic materials

- Master files:
  - Bit depth: 8-bit grayscale or 24-bit color.
  - Spatial resolution (photographs): 3000–5000 pixels on the long dimension (300–800 ppi).
  - Spatial resolution (maps and other graphic materials): 3,000 ppi along the long dimension.
  - Spatial dimensions: 100% of the original.
- Access and thumbnail copies use the same constraints as the test-based materials.
  - Access copies may be constrained to 150 ppi and 600 pixels across the long dimension.
  - Thumbnails may be constrained to 72 or 96 ppi and 150–200 pixels along the long dimension.

File Size Estimates

File size (in bytes) =
[(source width x height in inches) x (bit depth) x (pixels per inch)^2]/8.

A 5x7 inch, color photograph scanned at 24-bit, 600 ppi will create a 36 MB TIFF. *Note: 1 megabyte = 1,048,576 bytes, 1 gigabyte = 1,024 megabytes.*

>>>Please also see RESOURCES for links to tutorials and best practices.

Audiovisual Media

Audiovisual reformatting (or transfer) is much more complicated than image and text-based reformatting. Standards are evolving and largely based on increasing storage capability for the extremely large file sizes. There are very few non-proprietary and uncompressed or lossless options in file format or delivery. Minimum components required to complete the process include a source playback device, an analog to digital converter (ADC), a computer to process the digital signal, digital storage media, and editing and compression software. A good time estimate is four hours of labor for every one hour of media, especially when transcripts or captioning is required.

Special considerations for audiovisual media include:

- "Published" works have variant forms (theater version vs. director’s cut).

- "Unpublished" and ancillary materials: unedited footage, screen tests, scripts, storyboards, album covers, liner notes, lyrics, promotional materials, etc. (contextual materials)
- Many physical formats with limited life spans.
- Obsolescence of appropriate play-back equipment and technical expertise.
- Increased quantity of technical information to retain or obtain: such as frequencies, speeds, luminescence, tone references, channels, etc.
- Predictable loss of quality during reformatting process.
- Quality options (faithful, sufficient or informational/reference), file size and delivery method.

There are three primary delivery options for digital audiovisual media: storage media, download or streaming. **Storage media** may present certain restrictions on the file format, size or resolution, but may be portable (as with CD, DVD, etc). When providing files for **download**, the user must wait for the entire file to transfer; the user can use the file in any way they choose (presenting possible copyright issues); the user can search anywhere within the file; and the quality depends on what file size is available and how long the user is willing to wait for the file to transfer. With **streaming** delivery, the file can be viewed/listened to once the first part has been transferred; searching requires "communication" with the server and will have a delay; the user does not receive a complete copy; the quality depends on the network bandwidth. Additionally, a specialized server is required to allow streaming media, and the user must have a streaming media player (such as QuickTime, RealPlayer or Windows Media Player).

Digitization Priorities for Audiovisual and Electronic Media

According to Independent Media Arts Preservation, once an electronic media collection has been cataloged and the condition of its contents has been assessed, it is time to prioritize which material should be restored first. The highest restoration priorities are unique copies. If only one copy of a work exists, there is no recourse if it is lost or damaged. Other high priorities are works older than 15 years, if the content is valuable. Obviously, deteriorating or damaged material should also be given precedence.

A work's format can also help determine if it is a preservation priority. As a rule of thumb, three videotape formats require immediate attention: 1/2-inch open-reel videotape (most date from 1965 to 1975); 3/4-inch U-matic videocassettes (1973 to 1983); and original VHS tapes that are older than ten years with valuable content. Obsolete formats, such as 2-inch Quad video reels (1956 to 1981), also warrant immediate attention.

Audiotape formats that deserve high priority include: paper- and acetate-based 1/4-inch reels (from the 1940s and '50s); any recordings made between 1977 and 1983 on Ampex 1/4-inch audiotape; and unique audiocassettes made before 1990. For film, in addition to deteriorating and damaged stock, color-faded prints are a priority. For digital media, preservation is urgent if the software or hardware required to run a piece are threatened with obsolescence, or if the optical medium on which it is stored–CD-R, CD-ROM, or DVD–has begun to deteriorate. (IMAP, 2007).

>>>Please also see RESOURCES for links to more information and best practices.

Audio Media

Audio media may range from wire recordings and wax cylinders, shellac or vinyl disks (LPs), to magnetic tapes (1/4 inch reel-to-reel, 8-track, cassette tapes, etc.), including a variety of digital disk and tape formats. Like all traditional media, they are not without long-term preservation issues, including the availability of appropriate playback equipment. If stored properly, LPs, 78 rpms, reel-to-reel younger than 15 years, CDs and CD-Rs do not need immediate transfer. Technical expertise is often necessary to address play-back issues such as playback speed (earlier recordings often have inconsistent calibration), equalization, cleaning and conservation practices, and engineering problems using the obsolete equipment.

While analog media is able to capture sound as it is in nature, digitized sound is created by converting an audio wave to an electrical current to digital information, relying on sampling. When audio is digitized, the electric current is sampled at specific time intervals. The amplitude is recorded as a number at each sampling point. The sample rate and the bit depth determine the quality and resolution of digitized audio. The highest frequency pitch that a digital audio sample can record is one-half of the sampling rate: 44.1 kHz audio CDs can only record frequencies up to 22.05 kHz (CDP Digital Audio Working Group, 2006).

High resolution digital audio files are extremely large and practical storage or distribution methods rely on compression. Audio compression is lossy and based on psychoacoustics (based on understanding of human hearing and audio processing). In addition, audio compression is often propriety using a codec (a compressor/decompressor) such as RealAudio, Windows Media Audio or QuickTime, where the files can only be used with the respective players.

In 2002 presentation about audio transfer projects, George Blood strongly warned about the lack of standards for high-resolution digital audio – across all sectors, professional, archival or consumer. Three years later, he describes the emerging standards and issues a new caution:

> "What has changed in three years…? A place to keep the 96kHz/24bit or 96/24 files: permanent on-line storage, hard drives. Now we're not talkin' about $200.00 Firewire or USB drives. We're talkin' $10,000.00 enterprise-level storage with built-in RAIDs, drive and power supply backups, and a team of IT professionals to run backups and keep it all running. The kind of storage used by banks and VISA and where "the university" keeps transcripts and bursars records. At large institutions, these facilities are already available. You just have to make friends with the IT folks to store massive files (96/24 files are 1.1GB per track per hour!). For smaller institutions it may make sense to either skip the 96/24 files and work only with multiple CDRs, and/or CDRs in different formats (-DA and -ROM), or to use a third party to store the files. Just as you may use an outside vendor to hold paper records, there are vendors (such as OCLC) who keep archival files, providing the enterprise-level storage and IT support which are beyond the reach of a small institution (Blood, 2005)."

File formats

WAVE (.wav) is the "defacto" standard for master files with the Broadcast Wave (.bwf) becoming more popular. AIFF is another common format. **MP3** files are commonly used for compressed files.

Quality standards
It is important to keep in mind that audio digital standards are still evolving. As in text/image reformatting, the final product will also depend on the original. For example, it might be more appropriate to use higher quality for musical scores and a lower quality for spoken word or oral history. The "emerging standard" follows:

Master file: 96 kHz, 24-bit, WAVE or Broadcast wave (.wav or .bwf).
Access file: a 44.1 kHz, 16-bit, typically on a CD-DA (digital audio) disc
Web accessible format: such as .mp3, RA, or .qt (RealAudio or QuickTime)

File Size Estimates
Calculators are available online. 96 kHz, 24-bit files are approximately 1.1 GB per track per hour.

Moving Image Media

Analog moving image media collections may include nitrate or acetate films, disk (laserdisc, etc), magnetic/video tape (VHS, Beta) and digital tapes and disks. Properly stored, most film is reasonably stable and doesn't require immediate transfer (with a few exceptions such as nitrate film), but tapes are unstable as the physical components decay relatively quickly. Technical expertise and expensive (and obsolete) equipment are often required, and media can sometimes be difficult to identify. Unfortunately, most moving image media is very unstable and will need to be replaced sooner than later (none of the formats approach the stability of rag paper).

Digital video uses frame resolution and frame rate in addition to color space (luminance and chrominance or chroma) and audio. Digital video sampling is stated as a sampling ratio of the three parts of a component color difference (one luminance channel and two chroma channels); typical samplings are 4:4:4, 4:2:2, 4:2:0 and 4:1:1. For example, with 4:2:2, for every 4 samples of the luminance channel there are 2 samples for each of the chroma channels.

Like digital audio, high-resolution digital video files are extremely large and practical storage or distribution methods rely on compression. Compression is lossy, based on sampling, and often propriety using a codec (a compressor/decompressor) such as QuickTime. Formats may be selected based on the desired final product. What quality of sound or color fidelity is required? What additional features will be included: soundtrack, commentary track, captioning (or a text transcript), etc.?

File formats
Like audio, formats are still developing and standards emerging. **MPEG** is non-proprietary, but presents limitations (such as compression). **AVI** and **QuickTime** (with extensions) are proprietary, but incorporate more features such as support for multiple soundtracks or sound channels and captioning.

Quality standards
As the sampling rate increases, so the quality increases. In 4:4:4, the chroma channels are sampled equally to the luminance channel, creating better color definition, but an extremely high

file is created. Decisions will be made considering both the quality of playback experience intended (theater experience?) and available storage space and delivery options.

File Size Estimates
File size (in bits per second) =
Step 1: (pixel width) x (pixel length) = pixels per frame
Step 2: (pixels per frame) x (bits per sample) x (sample per pixel) = bits per frame
Step 3: (bits per frame) x (frames per second) = bits per second

Examples using 10 bits/sample and 3 samples per pixel:
4:4:4 sampling (720 x 486 resolution) **~141.58 gb/hour**
4:4:4 sampling (1920x1080 HDTV) ~ **840 gb/hour**
4:2:4 drops the sampling rate by a third
4:2:0 and 4:1:1 drops in half ~ **70 gb/hour** for normal resolution video

Born Digital Media

"Born digital" materials were created in a digital environment and have a much higher risk of being lost or altered, either making them unavailable as viable historic resources. The files are easily destroyed (or deleted) and the formats become obsolete. Born-digital materials include websites, e-mail correspondence, maps (and other information) generated from geographic information systems (GIS), all forms of computer files, digital video and audio recordings, digital photographs, etc. According to the Library of Congress, 44% of the websites online in 1998 had vanished one year later.

As with digital objects created by reformatting, long-term planning is required for their continued availability. A digital preservation plan will address rights management, migration to appropriate storage devices, best practices for creation and refreshing of digital objects, repository and backup practices, file-naming conventions and encoding ("tagging"). As in reformatting, use uncompressed, non-proprietary file formats when at all possible. Ideally, file formats will not be bound to a particular carrier (whether software or hardware/storage device) and will not be constrained, by passwords or other usage restrictions, to prevent backup methods and disaster recovery operations.

OUTSOURCING

Projects that require a high level of expertise, expensive or obsolete equipment, or a large time commitment are often ideal candidates for outsourcing. Digitization services for general business purposes often do not meet the preservation standards required by cultural institutions. It is a good practice to research digitization providers, speak to colleagues and contact institutions that have outsourced similar projects. The following guidelines are recommended (Blood, 2005):

1. ALWAYS check references.
2. There is NEVER one right answer to your questions.
3. Talk to your colleagues.
4. Check their qualifications.
5. Visit the vendor, or find someone who has.
6. Don't EVER talk to just one vendor.
7. Careful about assumptions of other lines of work the vendor does.

Audio reformatting is one area in which the engineering and conservation is quite complicated, the variety of media and play-back equipment is enormous, and reformatting is labor-intensive. The Association for Recorded Sound Collections (ARSC), has compiled a directory of members who offer preservation and restoration services as well as non-members who provide supplies and equipment (http://www.arsc-audio.org/audiopreservation.html). Independent Media Arts Preservation has also compiled a list of vendors experienced in transferring archival audiovisual materials (http://www.imappreserve.org/pdfs/Educate_Train_pdfs/Vendors.pdf).

>>>Please also see RESOURCES for links to more information and best practices.

FUNDING RESOURCES

Collections digitization is an expensive undertaking. Fortunately, there are an increasing number of funding opportunities for digitization-related projects and activities available from all levels: national, state, regional and private (via foundation or individual). It is imperative to know the scope and appeal of your collections and be aware of opportunities for collaboration. Following are a few examples of available funding opportunities.

Resources include the Getty Grant Program (http://www.getty.edu/grants/index.html), the Mellon Foundation (http://www.mellon.org/grant_programs/programs) and Technology Grant News (http://www.technologygrantnews.com/), which identifies public and private sector funding sources in the technology arena (published quarterly with a sample issue online).

Institute of Museum and Library Services (IMLS)
Main website: http://www.imls.gov/
Grant opportunities: http://www.imls.gov/applicants/applicants.shtm

Project types include collections management, community engagement, conservation, demonstration, digital collections/tools, formal education, informal learning, partnerships, public programs and research. Funds can be used for a wide variety of projects, including ongoing work, research and other behind-the-scenes activities, planning, new programs, purchase of equipment or services, activities that will support the efforts to upgrade and integrate new technologies, increase access to underserved communities, and create linkages between institutions. Projects should work to generate strategic impact, innovation and collaboration.

National Endowment for the Humanities (NEH)
Main website: http://www.neh.gov/
Grant opportunities: http://www.neh.gov/grants/
NEH Office of Digital Humanities http://www.neh.gov/odh/
NEH grants “of particular interest”:
http://www.neh.gov/ODH/GrantOpportunities/tabid/57/Default.aspx

The NEH provides many humanities projects related to digitization and the use of digital technologies. Opportunities include funding for start-up activities, creating tools and data analysis, creating and maintaining digital archives, collaborative projects, development of digitized teaching resources, public programming, and long-term support and improvements.

Activities may include digitizing collections, cataloging collections, preservation reformatting, preserving and improving access to "born-digital" objects, creating digital tools, databases and electronic archives.

National Historic Publication and Records Commission (NHPRC)
Main website: http://www.archives.gov/nhprc/
Grant opportunities: http://www.archives.gov/nhprc/announcement/

The NHPRC "promotes the preservation and use of America's documentary heritage essential to understanding our democracy, history, and culture." Gant opportunities focus on nationally-significant collections and include projects using cost-effective methods to digitize historical records collections, leading to sustainable electronic records archives and digital records preservation, developing new strategies and tools that can improve the preservation, public discovery, or use of historical records, basic archival activities promoting preservation and use of documentary heritage, detailed processing and preservation of archival collections, publishing of historical records, and strengthening state programs and building a national archival network.

National Center for Preservation Technology and Training (NCPTT)
Main website: http://www.ncptt.nps.gov/
Grant opportunities: http://www.ncptt.nps.gov/Grants/Default.aspx

NCPTT "protects America's historic legacy by equipping preservation professionals with progressive technology-based research and training. NCPTT seeks innovative projects that advance the application of science and technology to historic preservation. PTT Grants fund projects that develop new technologies or adapt existing technologies to preserve cultural resources." Project may include training activities, workshops, and curriculum development that promote the use of new or adaptive technologies; documentation using new or emerging methods; manuscript or website development that disseminates innovative preservation technologies; and meetings of experts to discuss the application of technologies to address preservation problems.

DIGITIZATION TERMS

Analog to digital conversion is the conversion of physical formats to electronic/digital objects. Pixel-based images (such as scanned photographs) are "raster" images or "bitmaps." In audiovisual reformatting, an ADC device (analog to digital converter) is used to link the analog play-back machine (such as a VCR) to the computer to translate the image/sound to a digital format.

Bit depth refers to the number of binary digits devoted to color information. Scanners generally have sensors to red, green and blue; each "channel" is stored separately in a digital file using 8 bits for each channel = 24 bits.

1 bit (1) = 2^1 = 2 shades (black and white) 4 bit (0010) = 2^4 = 16 shades
2 bit (01) = 2^2 = 4 shades 8 bit (11010001) = 2^8 = 256 shades

In audio digitization, bit depth is the range of numbers used to record each measurement.

Chrominance (“chroma”) is the color, hue and saturation information, indicates the difference between color components, represented as “CbCr” (for blue and red).

Codec is a device or program to compress and decompress a digital signal. It is used for audiovisual file formats to improve transmission of files across a network. CODECs are often proprietary (requiring specific player software, such as Windows Media Player) and use lossy compression.

Compression reduces the file size for storage or transmission. The files must undergo decompression to view. Compressed files should ideally only ever be considered for derivatives.

- Lossless compression allows reconstruction of the original information.
- Lossy is “visually lossless.” These files compress every time they are saved, discarding original information each time.

Derivatives are files created from the master; any image processing (color balance, cropping, etc.) is typically done on the derivatives. The quality and file type depends on the use. For example, low resolution and low quality thumbnails may be created for reference images in an online catalog or a “zoomable” format may be created for scholarly uses.

Frequency is number of times the wavelength occurs in one second. Measured in kilohertz (kHz), or cycles per second. The faster the sound source vibrates, the higher the frequency.

Luminance is the brightness, from white to black, represented as “Y”.

Master file refers to the high-quality original scan created from the earliest possible generation of the original. Ideally, the file will not use compression; will be retained for long-term storage; and will contain the most detailed information necessary for its purpose (meeting all foreseeable needs). Image processing on the master file is generally avoided, but color bars and rulers are often included for future reference. Make as many adjustments as possible (color, etc.) in the scanning software (before the final scan is complete) and use as much tonal range as possible.

Proprietary formats are “closed” and specific to a company or other limited group; they are often created for use with a specific program (.doc = Microsoft Word). These formats risk obsolescence and lack of support and should be avoided when possible.

Resolution is often referred to as dpi, dots per inch, or ppi, pixels per inch (“ppi” is the most technically accurate). It is the number of pixels, or dots, captured per inch of original object. For example, an 8x10 photograph scanned at 300 ppi = 2400 x 3000 pixels. The **spatial resolution** is the pixel dimensions of the image, 2400 x 3000.

Reformatting is converting or transferring one media type into another.

Sampling is the number of times per second the amplitude of the wave is measured. Sampling rate is described in kHz (thousands of samples per second).

RESOURCES / SOURCES

As stated above, there is a wealth of information available online. Many institutions, professional and governing organizations and repositories publish best practices, frameworks and guidance, decision matrices and much more detailed information on everything described above. In addition, there are many tutorials and webliographies available.

Following are a just a few examples, most of which cover all aspects of a digitization project or initiative (not only the specific topic). All links were verified as of June 18, 2008.

WEBSITES:

- BCR Collaborative Digitization Program, http://www.bcr.org/cdp/index.html
- Cornell University, Digital Imaging Tutorial, http://www.library.cornell.edu/preservation/tutorial/
- Conservation OnLine, Conservation Topics, http://palimpsest.stanford.edu/
- Digital Library Federation, Digital Preservation, http://www.diglib.org/preserve.htm
- The Getty, Standards and Guidelines, http://www.getty.edu/research/conducting_research/standards/
- Independent Media Arts Preservation, http://www.imappreserve.org/index.html
- Introduction to Imaging (Getty), http://www.getty.edu/research/conducting_research/standards/introimages/index.html
- Library of Congress, Digital Preservation, http://www.digitalpreservation.gov/
- Minnesota Digital Library, Digitizing and Metadata Standards, Best Practices and Policies. http://www.mndigital.org/digitizing/
- The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials http://www.nyu.edu/its/humanities/ninchguide/index.html
- Northeast Document Conservation Center, www.nedcc.org
- Northeast Document Conservation Center, Handbook for Digital Projects, http://www.nedcc.org/oldnedccsite/digital/tofc.htm

General:

Building and Sustaining Digital Collections: Models for Libraries and Museums. Council on Library and Information Resources. August 2001. PDF version, www.clir.org/pubs/reports/pub100/pub100.pdf

Besser, Howard. *Introduction to Imaging, Revised Edition.* The J. Paul Getty Trust. Online edition, http://www.getty.edu/research/conducting_research/standards/introimages/index.html

Chapman, Stephen. "Techniques for Creating Sustainable Digital Collections." *Library Technology Reports* 40(5), Sept./Oct. 2004. Online bookstore: https://publications.techsource.ala.org/bookstore/displayItem.pl?itemID=2542

Collaborative Digitization Program. *Getting Started with Digital Projects.* BCR. 2008. http://www.bcr.org/cdp/digitaltb/getting_started/faq.html

Gertz, Janet. *NEDCC Preservation Leaflet 6.6 Preservation and Selection for Digitization "New 2007."* Northeast Document Conservation Center. 2007. http://www.nedcc.org/resources/leaflets/6Reformatting/06PreservationAndSelection.php#top

Gonzalez, Maria E. *Preservation Reformatting Conference: Digital Technology vs. Analog Technology.*, Volume 26, Number 5. Abby Newsletter. July 2003. http://171.67.143.176/byorg/abbey/an/an26/an26-5/an26-502.html

Hazen, Dan, Jeffrey Horrell, and Jan Merrill-Oldham. *Selecting Research Collections for Digitization*. Council on Library and Information Resources, August 1998. http://www.clir.org/pubs/reports/hazen/pub74.html

Hazen, Dan, Jeffrey Horrell, and Jan Merrill-Oldham. *Selecting Research Collections for Digitization: a Decision-Making Matrix*. Council on Library and Information Resources, August 1998. http://www.clir.org/pubs/reports/hazen/matrix.html

Lavoie, Brian F. *The Incentives to Preserve Digital Materials: Roles, Scenarios, and Economic Decision-Making.* OCLC. 2003. PDF version, http://www.oclc.org/research/projects/digipres/incentives-dp.pdf

Lee, Stuart D. *Scoping the Future of the University of Oxford's Digital Library Collections, Appendix B: Decision Matrix for Proposed Digitization Projects.* University of Oxford. 1999. http://www.bodley.ox.ac.uk/scoping/report.html

Moving Theory into Practice: Digital Imaging Tutorial. Cornell University Library Department of Preservation and Conservation. 2000-2003. http://www.library.cornell.edu/preservation/tutorial/.

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials. Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, and the National Initiative for a Networked Cultural Heritage (NINCH), 2002. http://www.nyu.edu/its/humanities/ninchguide/index.html

NISO Framework Advisory Group. *A Framework of Guidance for Building Good Digital Collections.* 3nd edition. National Information Standards Organization. 2007. PDF version, http://www.niso.org/publications/rp/framework3.pdf.

Phased Custody and Delivery of Digitally-Reformatted Resources. Library of Congress. 2006. http://www.loc.gov/preserv/prd/presdig/presphase.html

Policy for Preservation of Digital Resources. Columbia University Libraries. July 2000. PDF version, http://www.columbia.edu/cu/lweb/data/services/preservation/dlpolicy.pdf

Preservation and Reformatting Section (PARS). *Definitions of Digital Preservation.* American Library Association. 2008. http://www.ala.org/ala/alcts/newslinks/digipres/index.cfm

Selection Criteria for Digitization. University of California Libraries. 2004. http://libraries.universityofcalifornia.edu/cdc/pag/digselec.html

Selection Criteria for Preservation Digital Reformatting. Library of Congress. 2006. http://www.loc.gov/preserv/prd/presdig/presselection.html

Sitts, Maxine K. (ed.) *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Northeast Document Conservation Center. 2000. PDF version, http://www.nedcc.org/oldnedccsite/digital/dman.pdf

Smith, Abby. *Strategies for Building Digitized Collections.* Council on Library and Information Resources, September 2001. PDF version, http://www.clir.org/pubs/reports/pub101/pub101.pdf

Smith, Abby. *Why Digitize?* Council on Library and Information Resources, 1999. http://www.clir.org/pubs/reports/pub80-smith/pub80.html

Sustainability of Digital Formats, Planning for Library of Congress Collections. Digital Preservation (Library of Congress). 2007. http://www.digitalpreservation.gov/formats/

Zorich, Diane. *A Survey of Digital Cultural Heritage Initiatives and Their Sustainability Concerns*. Council on Library and Information Resources. June 2003. PDF version, http://www.clir.org/pubs/reports/pub118/pub118.pdf

Rights Management:

Copyright Issues for Digital Collections, New Jersey Digital Highway. 2007. http://www.njdigitalhighway.org/copyright_issues_libr.php

Flowchart for determining when U.S. copyrights in fixed works expire. Bromberg & Sunstein, LLP. 2002. http://www.bromsun.com/practices/copyright-portfolio-development/flowchart.htm

A Guide to Deed of Gifts. Society of American Archivists. 1998. http://www.archivists.org/catalog/deed_of_gift.asp

Harper, Georgia. *Crash Course in Copyright.* University of Texas, 2001. http://www.utsystem.edu/OGC/IntellectualProperty/cprtindx.htm

Hirtle, Peter. *Copyright Term and the Public Domain in the United States* (current as of 1 January 2008). Cornell University. 2008. http://www.copyright.cornell.edu/public_domain/

InfoPeople. *Copyright Permission Agreement* [sample]. 2005. http://www.infopeople.org/training/past/2005/dig-copyright/CDLPermission.doc

Minow, Mary. *Library Digitization Projects and Copyright.* LLRX (Law Library Resource Xchange). June 2002. http://www.llrx.com/features/digitization.htm

Digital Asset Management (DAMS) and Repositories:

Current Practices in Digital Asset Management, Ver. 0.9. The Internet2/Coalition for Networked Information (CNI), Performance Archive & Retrieval Working Group. October 2003. PDF version, http://www.internet2.edu/arts/files/digital-asset-management(v09).pdf

Trusted Digital Repositories: Attributes and Responsibilities. OCLC/RLG. 2002. PDF version, www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf.

Waibel , Günter. *Special Issue: Managing Digital Assets in US Museums.* RLG DigiNews, Volume 10, Number 6. 2006. http://www.rlg.org/en/page.php?Page_ID=20999

Metadata and encoding:

Baca, Murtha, ed. *Introduction to Metadata: Pathways to Digital Information.* The J. Paul Getty Trust. Online Edition 2.1.
http://www.getty.edu/research/conducting_research/standards/intrometadata/

Guidelines for Electronic Text Encoding and Interchange. TEI (Text Encoding Initiative). 2007. http://www.tei-c.org/Guidelines/P5/

Image and text-based media:

Frey, Franziska and James Reilly. *Digital Imaging for Photographic Collections, Foundations for Technical Standards.*2nd Edition. Image Permanence Institute. 2006. PDF version, http://www.imagepermanenceinstitute.org/shtml_sub/digibook.pdf

Proposed Digital Imaging Standards and Best Practices, Indiana Memory and LSTA Digitization Projects. Indiana State Library. 2007. PDF version, www.in.gov/library/files/dig_imgst.pdf

Western States Digital Imaging Working Group. *Digital Imaging Best Practices.* Version 1.0. Western States Digital Standards Group (now CDP/BCR). 2003. PDF version, http://www.bcr.org/cdp/best/digital-imaging-bp.pdf

Audiovisual:

Besser, Howard. *Digital Preservation of Moving Image Material.* UCLA School of Education & Information Studies. 2001. http://www.gseis.ucla.edu/~howard/Papers/amia-longevity.html

Besser, Howard, et al. *Moving Image Preservation: Analog or Digital?* 18th Annual Preservation Conference: Preservation Reformatting: Digital Technology vs. Analog Technology. US National Archives & Records Administration. 2003.
http://www.archives.gov/preservation/conferences/papers-2003/besser.html

Blood, George. *Planning an Audio Preservation Transfer Project*, Presented at the convention of the Society of American Archivists, Birmingham, AL, Aug. 23, 2002, revised Jan. 12, 2005. Safe Sound Archive. 2005. PDF version,
www.safesoundarchive.com/PDF/AudioPreservProjectPlanning.pdf

Capturing Analog Sound for Digital Preservation: Report of a Roundtable Discussion of Best Practices for Transferring Analog Discs and Tape. Council on Library and Information

Resources and Library of Congress. 2006. PDF version, http://www.clir.org/pubs/reports/pub137/pub137.pdf

Casey, Mike and Bruce Gordon. *Sound Directions: Best Practices for Audio Preservation* Indiana University Digital Library Program. 2007. PDF version, http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/index.shtml

CDP Digital Audio Working Group. *Digital Audio Best Practices,* Version 2.1. BCR. October 2006, PDF version, http://www.bcr.org/cdp/best/digital-audio-bp.pdf

Lacinak, Chris. *Reformatting: Terminology, Intent and Practices.* Library of Congress, Moving Images Collections, 2005. http://mic.loc.gov/preservationists_portal/presv_reformtg.htm

An Overview of the Film Preservation Process. Library of Congress, Moving Image Collections. 2005. http://mic.loc.gov/preservationists_portal/presv_process.htm

Preservation 101. Independent Arts Media Preservation (IMAP). 2007. http://www.imappreserve.org/pres_101/index.html#priorities

Outsourcing:
Dale, Robin L. *NEDCC Preservation Leaflet 6.7 Outsourcing and Vendor Relation.* Northeast Document Conservation Center. 2007. http://www.nedcc.org/resources/leaflets/6Reformatting/07OutsourcingAndVendorRelations.php

***This leaflet was written for the Midwest Art Conservation Center in Minneapolis, Minnesota by Rebecca Holte, June 17, 2008. Thanks to Megan Winget, upon whose course much of this paper is built (UT Austin, School of Information, Survey of Digitization, Spring 2008).*