Guide to Digital Imaging

Utah State Archives
June 2005

Introduction

The purpose of this document is to provide information about digital imaging or the scanning of records. This guide specifically refers to records that were originally in another format and are captured digitally, rather than those records that are born electronically. It will discuss benefits and disadvantages, costs and efficiencies, retention and disposition of records, standards for digitization, and definitions. Agencies should justify the implementation of an imaging system based on an analysis of their work processes and business needs balanced against costs. A new imaging system should yield marked improvements in productivity and efficiency or quality of service. Please contact your analyst with further questions about your imaging project.

Factors to Consider When Digitizing Records

Courtesy of the National Archives and Records Administration. [Online version on May 19, 2005
http://www.archives.gov/records_management/policy_and_guidance/frequently_asked_questions_imaged.html].

Before starting any imaging project, know the project's mission, users, priorities (speed, image quality, and quantity), and functional goals (reference, web use, publication, other). Additionally, assess staff expertise and availability (to scan, manage infrastructure, migrate data, and build metadata), and address content issues, such as physical condition, format, nature and attributes to be captured. Realize that costs include more than just the initial purchase of an imaging system. You may also incur migration costs if the information has to be retained for periods longer than five to ten years.

What are possible advantages and disadvantages of maintaining records as digital images?

Advantages include:

- Ability to use very high-density electronic storage media instead of paper.
- Shorter retrieval time when the images are well indexed.
- Multiple users and access levels are possible.
- Low shipping costs and ease of information dissemination.
- Ease of use of imaged copies of records in vital records and disaster recovery plans.
- Ready access to digitized records may assist organizations needing to retrieve information efficiently during litigation and discovery.
- Ease of making copies of the imaged records.
- No loss of digital image quality from generation to generation. Well-made copies and derivatives can be as good as the original images.

Disadvantages include:

- Digital images are not human-readable without computer equipment.

- Significant equipment costs, including hardware and software.
- Potential for hardware and software obsolescence. Generally, systems change every 18 months to 5 years, software changes every 2-3 years, and the life expectancy of media is relatively short.
- Indexing requirements may be more extensive than is required with other formats. Unless records are arranged in a logical sequence or clearly indexed, it may be difficult to identify a series or use groups of records as a series.
- Different types of scanners may be required to scan text, oversize items, photographic prints, slides, and other formats.
- Digital quality control and image and metadata capture and management are complex, time-consuming processes requiring expertise.
- Complex disposition and potential problems in implementing dispositions need to be addressed and can include the following:

    If records are stored without regard to retention periods on an individual disk or in an individual directory, each record must be selected for destruction or moved to off-line storage.

    When agencies use write-once-read-many (WORM) optical media, records should be grouped by like retention periods on individual disks or in individual directories.

Other Factors to Consider:

- **Volume of records**. Imaging is generally used for large volumes of records.
- **Reference use**. Imaging is most effective on highly referenced collections where a short retrieval time is important or where there are multiple users accessing the same records. Combined with effective indexing, imaging records can facilitate retrieval.
- **Relationship to records on other media**. Consider whether the records to be imaged have to be used with records on other media.
- **Records and information usage**. Consider how the information is used and how long the record is needed. Required retention periods are specified in records schedules.
- **Legal acceptability**. Following established procedures and maintaining the documentation of audit trails and other business practices will ensure that information is kept that may be needed to document record authenticity and reliability.
- **Ease of maintenance**. Balance storage costs and capacity with indexing, conversion, quality control, and migration costs.
- **Staffing requirements**. Increased imaging and indexing of records and quality control procedures may require additional staff training.
- **Work process and information flow**. Would imaging facilitate the work process? Considerations include how records are routed, how information is added to records or files, and when records (finals or drafts) need to be captured.
- **Document preparation**. Determine how much work needs to be done to make the files ready for imaging. Document preparation for voluminous files may be significant.
- **Quality control issues**. Procedures must be instituted both while preparing documents for imaging and while verifying and validating imaged information.

- **Condition of original records**. The condition of the records will affect their handling during imaging as well as the quality of the imaged record that can be produced. This will particularly be a factor for records that are damaged, faded, or oversized.
- **In-house operation versus contracting operations with a service bureau**.
- **Image requirements** (resolution, compression, headers, etc.) will vary depending on how images will be used and the condition of the originals.
- **Indexing requirements** and metadata fields are determined by analyzing how users will access images.

Digital Imaging Cost-Benefit Analysis

Portions reprinted courtesy of Steven Puglia, "RLG DigiNews" [Online version on May 19, 2005 http://www.rlg.org/legacy/preserv/diginews/diginews3-5.html#feature].

Agencies initiate digital imaging projects primarily to realize increased efficiency and productivity through ready access to documents and information. Before any commitments to purchase equipment and software programs are made, a business plan should be developed which would specify the cost-benefits of installing such a system. Long-term preservation of the records should not be a major consideration when purchasing a scanning system. Reduced handling of materials is the only preservation-related benefit of digital imaging.

In developing a cost-benefit business plan you should consider the following:

Efficiencies and productivity increases

Efficiencies are gained primarily in saving of staff and/or client time. This could be quantified by calculating the staff hours saved and projecting those savings against the cost of the scanning system.

As an example:

- The time saved in finding, compiling, and retrieving files and information.
- Time saved in researching information that may be located in several different offices or regional centers.
- Cost savings through reductions of mailings and phone calls between offices.
- Some intangible benefits would result in a better quality of work product because of access to more reliable information.
- Clients would be served faster and provided with more complete information.

Cost considerations

- Match the capacity of the equipment to the overall size of the project. Do not overbuy but assure the equipment is scalable to accommodate any new or expanded projects. Future needs of the agency should be considered to determine if the scanning project will expand into an enterprise system.

- Take a representative sample of the documents and develop procedures for document preparation. This will include arrangement and removing extraneous material such as staples, etc. From this review, a determination can be made as to time and costs involved.
- Scanning time can also be determined by using a sample of documents and actually scanning them into a similar system. The vendor should be able to provide for this review.
- The description, indexing, cataloguing, and development of finding aids for the material is very time consuming and costly. This should be well-planned out to achieve the desired end results of efficiencies or added value to the information system.
- Network infrastructure, computer storage space, and maintenance contract costs should be spelled out and included in the overall costs of the system.
- The eventual cost of migration of hardware and software to new systems should be taken into consideration. Eventually this will need to be done to upgrade and maintain the integrity of the system. These costs should be identified by the vendor and built into the overall costs of the system.
- Cost analysis studies of existing imaging systems are showing the following cost trends: one-third of the overall cost is attributable to digitizing (equipment, document identification, preparation, and scanning); one-third to metadata creation (cataloging, description, indexing, finding aids); and one-third to system administration (system infrastructure, maintenances, quality control, migration).

The costs per image will vary widely depending on agency needs and the designs of the system. These costs may range from $1.85 to as much as $96.45 per image. See this cost study from 1999 in detail at: http://www.rlg.org/preserv/diginews/diginews3-5.html.

Retention and Disposition of Imaged Records

Portions reprinted from the National Archives and Records Administration. [Online version on May 19, 2005
http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf].

Like paper records, digital records must be appropriately scheduled and retentions linked to a general schedule item or approved by the State Records Committee. Before beginning a digitization project it is important to have the record series scheduled and the retention approved because the approved retention must be built into the project from the outset.

Digitized records may exist in more than one format. For example, there may be a paper copy as well as an electronic copy. If so, it is important to determine which format is the record copy (the copy to be maintained according to an approved retention schedule). If the agency determines that the electronic copy is the record copy and the record has a permanent retention, then the agency must be prepared to maintain the records indefinitely in an electronic format. This means that the agency must not only have storage space, but will also have a schedule for migration and/or conversion of the records as technology changes. Depending on how the record is used and by whom, one option might be to maintain the paper version as the permanent record and use electronic copies for office access.

For electronic records that are to be maintained permanently, the National Archives and Records Administration (NARA) recommends storage on hard drive systems with a level of data redundancy, such as Redundant Array of Independent Disks (RAID) drives, rather than on optical media, such as CD-R or DVD-R. An additional set of images with metadata stored on an open standard tape format (such as Linear Tape Open [LTO]) is recommended. (CD-R or DVD-R backup is the less desirable option.) Regular backup onto tape from the RAID drives is also recommended. A checksum should be generated and stored with the image files so that any data lost or altered in migration can be detected.

Optical disks are for distribution of images to external sources and not for long-term storage. If CD-Rs (such as Mitsui Gold Archive CD-Rs) or DVD-Rs are used they should be high quality. Two copies should be made and one should be stored off-site. Optical disks should be regularly checked for data integrity.

Compatibility and obsolescence in the future are likely to be problems. Optical disks are easily damaged by dust and fingerprints and are also especially susceptible to damage from excessive heat or humidity. Most importantly, optical disks are likely to become technologically degraded within 5 to 10 years. If optical disks are used to store records with retention periods that exceed their expected life span then it will be important to have a data migration or conversion policy and procedure in place for transferring these records to the next generation of hardware and software.

Digitization of archival records and creation of metadata represent a significant investment in terms of time and money. It is important to realize that the protection of these investments will require the active management of both the image files and the stored metadata. *Storing files to CD-R or DVD-R and putting them on a shelf will not ensure the long-term viability of the digital images or their continued accessibility.*

Digital preservation remains a challenging area in which techniques, costs, and skills are still in development. Institutions increasingly invest heavily in digital materials but policies and procedures for long-term management of digital assets remain underdeveloped.

For records that do not have a permanent retention, a means of identifying and destroying records must be built into the software at the outset of the digitizing project. Erasure should take place according to approved procedures established by the agency. Erasure must be documented in the same manner that destruction logs are kept for the destruction of paper records.

Standards for Digitization of Permanent Records

Portions reprinted from the Western States Digital Standards Group. [Online version on May 19, 2005 http://cdpheritage.org/westerntrails/standards.html].

Basic Principles

- Scan at the highest resolution appropriate

- Scan at an appropriate level of quality to avoid re-scanning
- Scan the original document rather than a copy in order to capture the best quality image
- Create an uncompressed master image file which can be used to produce access copies
- Select equipment based on optical resolution as opposed to interpolated resolution as this will produce more accurate scans
- Use non-proprietary components
- Monitor and recopy files as needed
- Implement a migration strategy to transfer data across generations of technology

  -Transfer files to new media as it becomes widely available
  -Do not let more than 5 years elapse before refreshing your data
  -Longevity is less important than the ability to access

Initializing the Project

The following standards represent the industry recommendations for permanent historical records. The specifications have been compiled from a number of sources noted at the end of this guide. These are general guidelines to assist agencies when planning digitization projects. They are considered the **minimum** specifications necessary to obtain an acceptable level of image quality and ensure the long-term preservation of the permanent records of the state.

Resolution: The number of pixels (in both height and width) making up an image. The more pixels in an image, the higher the resolution, and the higher the resolution of an image, the greater its clarity and definition (and the larger the file size). Resolution can also refer to the output device, such as a computer monitor or printer, used to display the image. Image file resolution is often expressed as a ratio (such as 640x480 pixels), as is monitor resolution; however, resolution is also expressed in terms of pixels per inch (ppi). Image file resolution and output (print or display) resolution combine to influence the clarity of a digital image when it is viewed.

Modes of Capture

- *1-bit or Bitonal*: Means a pixel can be black or white. Bitonal imaging is ideal for black and white images, such as line drawings and text. However, scanning in grayscale rather than bitonal may produce a better looking image.
- *8-bit grayscale*: Each pixel can be one of 256 shades of gray. Good for black and white photography, and handwritten documents.
- *Color*: Each pixel can be one of 16.8 million colors. Used for documents with continuous tone color information. Color images should be saved as RGB files instead of CMYK files, due to the limited color range of CMYK.

Permanent Records Standards for Digital Imaging

Material	Bit Depth	Resolution	File Format
Printed Text	1 bit bitonal	300-600 ppi	Uncompressed TIFF
Handwritten Text	8 bit grayscale	200-400 ppi	Uncompressed TIFF
Photograph (B&W)	8 bit grayscale	3000-6000 pixels across long dimension	Uncompressed TIFF
Photograph (Color)	24 bit color	3000-6000 pixels across long dimension	Uncompressed TIFF
Maps/Oversized	8 bit grayscale or 24 bit color	200-400 ppi	Uncompressed TIFF

Access/Format

It is recommended that agencies first create a master image of high quality and then derive other versions from the master for general access. Below are suggested file formats for master and derivative images.

Master Image (uncompressed & unedited to remain as similar to original as possible)

- **TIFF** (Tagged Image/Interchange File Format)

Standard format for archival images allowing for a higher level of detail. TIFFs can be compressed (lossless compression maintains the original picture quality) or uncompressed. TIFF is chosen as the format for master images due to its interoperability, large data capture, and non-proprietary nature.

Derivative Images (can alter image, resize it, and save it in format for easier viewing; usually involves a loss of information)

- **JPEG** (Joint Photographic Experts Group)

24-bit, lossy (some data is lost) compression format ideal for screen and print presentation. The original image is reduced and cannot be restored. The file sizes are much smaller than TIFF but JPEG is not recommended as an archival file format.

- **GIF** (Graphic Image File)

Widely supported image storage format; good for low resolution screen display of images such as thumbnails or screen versions of text documents. Lossy compression.

- **PNG** (Portable Network Graphics)

Designed to replace GIF due to PNG's smaller file size and lossless compression.

- **PDF** (Adobe)

Provides a convenient way to view and print images at high resolution.

Definitions

Bitonal: Means that each pixel can be either black or white.

Checksum: A program which allows the user to see how many bytes were present before and after migration to detect information lost during migration.

Compression: The reduction of image file size for processing, storage, and transmission. The quality of the image may be affected by the compression techniques used and the level of compression applied. There are two types of compression:

Lossless compression is a process that reduces the storage space needed for an image file without loss of data. If an image has undergone lossless compression, it will be identical to the image before it was compressed. Primarily used with bitonal images.

Lossy compression is another process that reduces the storage space needed for an image file, but it discards information that is "redundant" and not perceptible to the human eye. If an image that has undergone lossy compression is decompressed, it will differ from the image before it was compressed, even though the difference may be difficult for the human eye to detect.

Conversion: The task of moving data from an existing format to a different format due to obsolescence.

Derived Image (Derivative Image): An image that has been created from another image through some kind of automated process, usually involving a loss of information.

Dots per inch (dpi): A measurement of the scanning resolution of an image or the quality of an output device. DPI expresses the number of dots a printer can print per inch or that a monitor can display both horizontally and vertically.

Interpolate: To calculate or estimate intermediate values occurring between two known values. This function is acquired artificially through software.

Linear Tape Open (LTO): An "open format" technology, which give the user multiple sources of product and media to choose from.

Metadata: Data about data, or information known about the image in order to provide access to the image. Usually includes information about the intellectual content of the image, digital representation data, and security or rights management information.

Migration: Preserving the integrity of digital images by transferring them across hardware and software configurations and across subsequent generations of computer technology. Migration includes refreshment (copying digital files from one media to another) as a means of preservation and access. However, migration differs from refreshment in the sense that it is not always possible to make an exact copy of a database or even an image file (as changes in hardware and software occur) and still maintain compatibility with the new generation of technology.

Pixels: Often referred to as dot, as in "dots per inch". "Pixel" is short for picture elements, which make up an image, similar to grains in a photograph or dots in a half-tone. Each pixel can represent a number of different shades or colors, depending on how much storage space is allocated for it. Pixels per inch (ppi) is sometimes the preferred term, as it more accurately describes the digital image.

Redundant Array of Independent Disks (RAID): A redundancy subsystem of disk drives that improves performance, fault tolerance, and data recovery.

Resolution: The number of pixels (in both height and width) making up an image. The more pixels in an image, the higher the resolution, and the higher the resolution of an image, the greater its clarity and definition (and the larger the file size). Resolution is expressed in pixels for image files or as dots per square inch (dpi) for prints. It can also refer to the output device, such as a computer monitor or printer, used to display the image, which is expressed as dots per inch.

For additional terms please visit
http://www.cdpheritage.org/resource/introduction/rsrc_glossary.html#P

Bibliography

These URL's were valid on May 19, 2005.

Colorado Digitization Program:
www.cdpheritage.org/resource/scanning/documents/WSDIBP_v1.pdf

National Archives and Records Administration: Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images by Steven Puglia, Jeffrey Reed, and Erin Rhodes (June 2004)
http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf

National Archives and Records Administration: Expanding Acceptable Transfer Requirements: Transfer Instructions for Existing Permanent Electronic Records Scanned Images of Textual Records
http://www.archives.gov/records_management/initiatives/scanned_textual.html

National Archives and Records Administration:
http://www.archives.gov/records_management/policy_and_guidance/frequently_asked_questions_imaged.html

Standards and Guidelines for Digitization and Scanning: OhioLINK Digitial Media Center
http://www.ohiolink.edu/media/dmcinfo/ScanningStandards.pdf

Steven Puglia, "RLG DigiNews" http://www.rlg.org/legacy/preserv/diginews/diginews3-5.html#feature

Steven Puglia, "Test-Driving the Technology," Presentation delivered at the conference of the Northeast Document Conservation Center held in Chicago, IL (June 3, 2004).

University of Illinois Photo Image Quality Calculator:
http://images.library.uiuc.edu/projects/calculator/

University of Tennessee Digital Library Center Digitization Standards and Procedures:
http://diglib.lib.utk.edu/dlc/techdocs/UT_DigitizationStandards2004.pdf

Western Trails Digitization Best Practices: http://cdpheritage.org/westerntrails/standards.html

For additional information please visit the Utah State Archives website at www.archives.utah.gov.