Kentucky Department for Libraries and Archives
Public Records Division

Ensuring Long-term Accessibility and Usability of Textual Records Stored as Digital Images: Guidelines for State and Local Government Officials

Revised (January 2010)

Introduction

Ensuring long-term accessibility and usability of textual records stored as digital images is largely dependent on the design, implementation, and management of the digital imaging system. A common misperception is that imaged records will be available as long as the physical media used to store the images are maintained. Long-term accessibility to records also requires system functionality between hardware platforms, software platforms, and storage media over time, as well as the ability to transport records and access tools. The life of an imaging system is conservatively estimated at about three years, while records retention and access requirements often exceed this short lifecycle. Appropriate policies, management procedures, and technology shall be applied from the design of a system until it is redesigned or migrated, to ensure that long-term records are accessible throughout their legal retention period. This requires a commitment of resources to preserve the accessibility and usability of digital images.

Legal and Policy Considerations

Specific laws and regulations related to government functions define how records are created, formatted, and maintained. These requirements, as well as legal minimum retention periods established by the State Archives and Records Commission in the form of records retention schedules, shall be identified when planning an imaging application. KRS 171.660 authorizes optical imaging as a valid duplication or reproduction media, if criteria required by the Kentucky Department for Libraries and Archives (KDLA) and 725 KAR 1:020 are met. Optical imaging can be challenged in court, based on inappropriate controls surrounding the creation and storage of images, so agencies shall be prepared to provide system documentation describing these controls to courts seeking authentication of legal documents.

Management Considerations

The following management considerations shall be addressed in system planning, development, and migration.

1.   System Planning, Acquisition, and Development

Budget for Change

Although there is an initial start-up cost for imaging hardware and software, the bulk of imaging costs cover training, conversion of documents to digital formats, indexing, and upgrades to current technology. On average, annual budgets should include from ten percent to twenty percent of the original cost of the system for maintenance, support, and upgrades.

Conduct a Requirements Analysis

Agencies shall conduct a requirements analysis before designing an imaging system. The Commonwealth Office of Technology (COT) provides guidance for technology project planning in the Commonwealth's Project Management Framework: http://technology.ky.gov/epmo/PMF_intro.htm.

Select a Reliable Vendor

State agencies shall select a vendor from state price contract. Local agencies are encouraged, although not required, to select from the state price contract. Imaging software is specified in the Kentucky Enterprise Architecture and Standards:

> 2850 - Imaging Solution - OCR/ICR Software
> 2860 - Imaging Solution - Workflow Software
> 2870 - Imaging Solution - Control Software for Imaging
> https://gotsource.ky.gov/docushare/dsweb/Get/Document-301104/

2. System Management

Effective management of an imaging system allows an agency to benefit fully from it and to address any associated risks. Proper management of an imaging system ensures long-term integrity and authenticity of imaged records and their admissibility in legal proceedings and audits. Special attention shall be given to any system that produces records for use in legal proceedings and audits, or a system that potentially exposes an agency to a high degree of risk. Records retention is a critical factor in managing an imaging system, as noted in KDLA's Policy Memorandum on the Storage of Public Records as Scanned Images (PM 2010-01).

3. Migration

When records have a lengthy retention requirement, a migration strategy is an essential component in ensuring long-term access to usable imaged records. Such a strategy shall provide for moving records from one generation of technology to another and for maintaining functionality of the records. Migration strategies shall take into account the use of open systems, standards-compliant technology, budgets that provide for training and technology upgrades, selection of a dependable vendor, and sound management of the system.

Plans for a technology strategy shall include continual actions such as:

- ensuring the preservation of imaged records on existing media through careful attention to environmental storage
- maintaining the functionality of existing hardware and software through upgrades of equipment and source code
- transferring the images, indexes, and other related data through successive versions of hardware and software
- migrating optical imaging systems to successive generations of technology, as yet undefined
- monitoring technology developments and trends and modifying migration plans as needed

Technical Considerations

1. Migration

The ability to migrate to new technologies is vital to retaining imaged records over a long period. The technology choices made when systems are developed or upgraded will often determine the options for future migrations. New technology shall comply with the following guidelines:

> **Select an open system solution.** An open system solution is one in which the hardware and software components are purchased from different vendors and integrated into a system. The open systems approach provides a maximum amount of choice to the system developer and end user of the system. Software used in an open system is "portable", which means that it can be moved to a variety of hardware. The software is also "scalable", which means that a system can be sized to handle both small volumes of users and records and can be expanded to larger

volumes. Open systems can therefore be scaled up with limited disruption to operations, including the maintenance of records.

Select standards-compliant system components. System components that are compliant with industry standards and best practices are easier to upgrade and migrate. Imaging hardware is specified in the Kentucky Enterprise Architecture and Standards:

> 1200 Scanners **-** Digital Imaging
> https://gotsource.ky.gov/docushare/dsweb/Get/Document-301102

Make controls and system auditing tools available. Systems shall be capable of providing audit trails and system security. Effective audit trails can automatically detect who had access to the system, whether staff followed existing procedures, or whether fraudulent or unauthorized acts occurred. Software is available for keystroke monitoring, time and date stamping, virus detection, and other controls that can be built into the design of systems.

Select appropriate storage media and environments. Information and images shall be stored on a server (or a mainframe acting as a server) and backed up either on a different computer or on different media. WORM (write once, read many times) technology is recommended for offline storage of imaged records when long-term retention and legal admissibility are the primary considerations. Many other media, however, may be suitable. If CD-ROM is used as a storage media, it shall comply with the ISO 9660, CDFS (Compact Disc File System) which specifies how a CD-ROM disk stores information. Regardless of media selected, government agencies shall not operate drive systems in environments with high levels of airborne particles and shall periodically clean optical media to remove dust and other particulates.

2. Document Preparation

Proper and thorough document preparation is vital to the success of a scanning program and is critical in achieving satisfactory throughput rates. In preparing textual documents to be scanned, agencies shall follow these minimum guidelines:

1. Documents shall be removed from folders, binders or other containers.
2. Documents shall be unfolded and stacked in the proper sequence.
3. Each record shall be verified for completeness.
4. Missing portions of a record shall be located and inserted into the proper place before the record is scanned.
5. All staples, clips, and other fasteners shall be removed.
6. All torn or damaged documents shall be repaired.
7. Proper scanning order shall be followed. In many cases, the type of record being scanned determines the scanning order. Transaction or batch documents shall be scanned in their sequential order or in the order in which they are received into the agency.
8. Case files shall be scanned together. This grouping of related documents or sequential ordering of like documents will reduce the amount of media interchange during document retrieval.

Regardless of the scanning sequence or the preparation method being used, the entire process shall be tested using sample records or documents. Some documents may present special difficulties in scanning and may require enhancement. Maps, charts, and other documents in which the scanner reads the foreground and background as the same color may not scan successfully.

3. Scanning Resolution

Scanners are available commercially with resolution capability between 100 x 100 and 600 x 600 dots per inch (dpi). For standard text documents, 200 dpi shall be the minimum resolution. Line breakup at 200 dpi can occur with 6 point or smaller typefaces. If the document has at least 9 point type or larger, 240 dpi or 300 shall be used. The quality of the text depends on the point size and the type of font being

scanned. Also, scanning at the lower dpi will result in the loss of any weak or fine line signatures. A gray scale level of at least 16 bits and a minimum of 300 dpi are recommended to aid optical character recognition.

For capturing detail in drawings or photographs, gray scale and/or a minimum of 300 dpi shall be used. Many scanning systems have prescribed photograph settings which control half-tones, gray scale sharpening, brightness, and contrast. Testing of prescribed and manual adjustments for photographs is necessary to get the best results. Color photographs with important details may require 24 bit (millions of colors) to accurately capture the image, though 8 bit (256 color) is adequate if images will be viewed exclusively on a computer monitor.

Since higher resolution or use of gray scale may result in larger files, select the resolution necessary to insure the quality of the record. If the original documents are retained and the scanning is for access rather than preservation, then the resolution requirements may be less stringent. This requires the image to be read easily on a display monitor or when printed out. If the originals are destroyed, a higher resolution shall be required, as there are very few options in restoring information after low resolution scanning.

Some compression algorithms (JPEG particularly) can result in loss of data, particularly if the compression is maximized above a 10:1 ratio. Photographic images and large paper documents, such as maps and drawings, may require use of different compression techniques with a measured compression ratio based on details to be preserved. The recommended practice for text documents is lossless compression. If higher resolutions from use of gray scale create a network traffic problem, the agency may consider use of a two part or two layer TIFF file to speed retrieval to the screen. The lower resolution image is used for screen display and the higher resolution image is used to store and print a quality image.

The following table shows examples of file sizes at various resolutions and use of gray scale.

File size of a standard 8 1/2 x 11 text document (compressed and uncompressed) at various dpi and use of gray scale.

	200 dpi	300 dpi	300 dpi 4 bit gray scale	300 dpi 8 bit gray scale	400 dpi
Uncompressed	500K	1.05 MB	4.08 MB	8.3 MB	2 MB
Group IV	50K	105 K	400K	800K	220 KB

4. Quality Control, Image Inspection, and Verification

Quality control refers to the methodology and techniques used to ensure consistency of procedures and output. Rigorous quality control procedures shall be used to ensure that the recorded images are of acceptable quality and can be accurately retrieved with the indexing method employed.

Image inspection verifies that the scanning procedures and equipment are producing a digitized image of acceptable quality. Immediately after scanning, the image is displayed on a monitor. The operator shall confirm that the document is legible, that no corners are folded or parts of the document otherwise obscured, that the document is right-reading, and that the image is of acceptable quality. During this inspection and verification sequence, the image shall reside on a magnetic buffer (a tape, cartridge, or internal hard disk). Following inspection, verification, and accurate indexing, the image shall be sent to the optical disk storage system.

Following recording of the image on optical disk, the image shall be retrieved through the system and verified against the original document to ensure acceptable recording quality. This procedure shall be performed daily during the first several weeks after system implementation, and periodically thereafter, until consistent recording quality is confirmed.

If images cannot be verified as completely readable and accurate after corrective rescanning, the corrective measures shall be documented, the image shall be identified as inaccurate, and the original hard copy document shall be kept. The location of the original shall also be part of the image documentation. If the scanner operator can identify information in the original document that does not appear clearly on the scanned image, this information shall be attached on a separate note with the date and operator's name recorded. Nothing shall be written on the original document; writing on the original invalidates the authenticity of the image copy as a valid duplicate of the original. Most digital imaging systems permit the attachment of a note to an image as a separate field in the image database. If this feature is not available on the system being used, then a note shall be created on paper, scanned as a separate image, and attached to the document.

Quality control testing of the scanner, through proper use of the technical targets and user targets described in ANSI/AIIM MS44-1988 (R1993), Recommended Practice for Quality Control of Image Scanners, shall be performed before and after scanning each batch of documents. This determines if the scanner is maintaining proper calibration during scanning and is operating within acceptable parameters. All testing shall be documented. The critical factor in determining acceptability of scanned images is the hard copy output of the images. Quality of images cannot be accurately determined by on-screen inspection. Image output shall be compared to a quality reference target output from the system to determine acceptability.

Training and supervision of operations staff is a key factor in maintaining acceptable image quality. There are no objective empirical indicators of acceptable image quality for digitally scanned images. An alternative is to categorize documents based on scanning problems and reach a consensus on how to most effectively capture the "best" image. Ideally, this decision process involves a team consisting of image system production staff, records managers, and system users and researchers. These evaluations shall include visual analysis of workstation display screen images and laser printer output. Retaining a set of representative laser prints for future reference is also a valuable image analysis benchmark tool.

5. Location of Index Database

Typically, image index databases for optical media systems are stored on separate magnetic storage media (usually a hard disk drive). For "fail-safe" storage of the database index, information may also be written to several locations of the optical disk(s) that the index information is derived from.

6. Defining Indexing Requirements

Indexing parameters are the categories of information by which document images are indexed for retrieval. The parameters shall be based on the retrieval needs of current and future users of the system. The selection of indexing parameters shall occur at the time the system is designed. This will allow the system designers to modify system components to accommodate indexing needs. Indexing parameters consist of fields that are used to locate and retrieve images. An index can also include fields which provide information about a record, such as a summary, notes, or other bibliographic information.

Logical and straightforward retrieval of records is dependent on appropriate indexing parameters. Agencies shall ensure that current and future access is considered when creating a system for records requiring long-term storage. Although indexing parameters can be added or amended in some programs, this process is costly and difficult to implement.

7. Index Data Entry

Digital imaging systems use databases to serve as indexes for the digital images stored on optical disks. A database record is created for each image. This record is broken into fields that correspond to the system's indexing parameters. Data entry shall be accomplished while scanning is done, immediately after, or at some longer interval. Data entry shall be done via key-entry, downloading of values, or auto indexing. Regardless of the data entry method chosen, databases are usually maintained on magnetic

media. Access and retrieval speeds for optical disks are significantly slower than magnetic media. The indexes can be written to their associated optical disks for security.

For more information or assistance on long-term maintenance of images, contact KDLA's Technology Analysis and Support Branch.

For information about KDLA's comprehensive document backfile conversion services for Kentucky public agencies, contact KDLA's Micrographics and Imaging Branch.

Glossary of Imaging Terms

Audit trails - A procedure or process that documents who used the system, when they used it, what they did while using the system, and the results.

Backfile conversion - The process of scanning, indexing, and storing a large backlog of documents on an imaging system.

Compression algorithm - A software or hardware process that "shrinks" images to occupy less storage space, and to provide for faster transmission.

Eye-readable - Images which can be read by the human eye. Paper and microforms are eye-readable. In the case of microforms, eye readability requires magnification.

Gray scale - The spectrum or range of shades of black an image has. Scanner and monitor gray scales are determined by the number of gray shades or steps they can recognize and reproduce. A scanner which can see a gray scale of 16 will not produce as accurate an image as one that distinguishes a gray scale of 256.

Half-tone - A graphic, usually created from a photograph, in which dots are used to represent the continuous tones that are in the original photograph. This is often expressed in lines per inch.

Image resolution - The fineness or coarseness of an image as it was digitized, measured as dots per inch (example 200 x 200 dpi often expressed as 200 dpi). The higher the resolution, the greater the amount of detail shown.

Optical character recognition (OCR) - The ability of a scanner with the proper software to capture, recognize, and translate printed alphanumeric characters into machine readable text. OCR uses either pattern matching or feature extraction. With pattern matching, the software has a template of possible characters. A letter is compared to a library of patterns.

Scanner - A device that optically senses a human readable image and contains software to convert the image to machine readable code.