Guidelines For Scanning University Records

Scanning, or digital imaging, is an increasingly popular strategy for dealing with records. Scanning can be a useful tool for managing your records and enhancing workflow, but is not always a good idea. Anyone thinking of scanning their records needs to keep a number of issues in mind.

The key distinction to make when considering scanning is between access and preservation. Scanning is not inexpensive, and it is not a good strategy for long-term preservation of records. It is almost never a good idea unless it is used to create better access to the records.

What is scanning?

Scanning, also referred to as digital imaging, is a process whereby a document is converted from print to a computer-readable format. You can think of the digitized version as a photocopy that can be viewed on your computer. Digital images produced by scanning are equivalent to the photographs one produces with digital cameras: they can be transmitted, displayed, and printed, but as images they are not text searchable. In order to make searchable electronic text, one must either transcribe records by typing or perform optical character recognition (OCR) processes upon digital images following scanning.

Why would I scan records?

Scanning's great strength is as a means of providing access to records. When records need to be accessed frequently, or from remote locations, or simultaneously by multiple users, scanning can be a cost effective means of distributing and rendering information. However, if full-text searching is required, the cost will go up considerably, due both to the OCR process itself and increased quality-checking.

Scanning records to save on storage costs is not likely to be cost-effective. Always do a full cost analysis before attempting this.

An analysis is also needed before scanning records that need to be retained for a long period of time, or which are to be retained permanently. The greatest issue with managing all electronic information is technological obsolescence. This means that the technology used to read scanned records is advancing at a great rate, and the systems needed to read your records may become obsolete long before your need for the records has ended. In this case you will need to plan - and budget - for periodic migrations of the records to newer systems. The digital images will also need to be reformatted as the software used to create and read the digital images becomes obsolete. Additionally, for records with permanent retention periods, the original paper documents may need to be maintained as well as the digital images.

What is the scanning process?

The following is a quick overview of the scanning process:

Document arrangement: Prior to scanning, determine the units of organization for the digital copies. Will they mimic the arrangement of the original prints, or will they, for example, be separated from one source item into multiple documents? Scanning is not always a one (source)-to-one (digital copy) process.

Document preparation: Physically clean up the documents to prepare them for scanning (remove staples, unfold paper, remove extraneous documents, etc.)

Identification: Consider what metadata (information about the documents) will need to be made to describe and organize the digital copies. This metadata may be recorded in something as simple as a file-naming convention or as complex as an indexing system. The primary reason for scanning is to facilitate access to the records. Batch-level scanning cannot automatically associate a group of digital images with a specific document or record. Plan for additional procedures and possibly systems to generate appropriate metadata to accompany digital images.

Technical considerations: Decide on file formats and other technical requirements for scanning, storage, and retrieval.

Quality control: Images must be inspected to ensure that they are of good enough quality for the purpose for which they are being scanned. In some cases, every image must be reviewed, in others only a sampling.

Storage: Digital files and digital media are inherently fragile. Regardless of storage media used, it is always prudent to make multiple copies and, ideally, to store the copies in separate locations-even during the production phase of a scanning project.

Disposition of source documents: Discard the paper once you are satisfied that the electronic records are accurate. The scanned images are generally an acceptable substitute for the original documents provided the scanning process has been carefully documented and the authenticity of the records ensured, and the images themselves are useable. Your scanning workflow must have safeguards and controls built into it so that you can assert to the satisfaction of a court that the images reproduced from the system are accurate representations of the original documents and that the information in the system has not been tampered with.

Migration, beyond storage: As discussed above, the greatest issue with scanning is technological obsolescence. A plan for forward migration of digitally imaged records must be put in place at the outset of the project and monitored as long as the records exist. Digital records are as bound by retention requirements as those in hard copy and failing to migrate records forward if they are still within their required retention period can hurt your office in the event of litigation or audit.

How much does scanning cost?

As you might guess from the above, a scanning project is an expensive undertaking. The actual scanning of the documents is the cheapest part of the process. Preparing the documents prior to being scanned can easily account for one-third of the project budget. And ensuring the records remain available over time may be even more costly. For paper records, it can take many years for the cost of scanning to catch up to the storage cost. The following table has cost estimates for just the scanning step (which might not be the most expensive part).

COST PER PAGE (SCANNING ONLY)	PAGES PER BOX	COST PER BOX	COST FOR ONE YEAR OF STORAGE	NUMBER OF YEARS TO BALANCE COSTS
$.10	3000	$300.00	$9.00	34
$.35	3000	$1050.00	$9.00	117

Since scanned records must be migrated forward with hardware and software changes, you will have to budget for this on an ongoing basis.

All of these numbers will vary, so any office considering scanning their records should do a full cost analysis.

What are the storage format and media issues?

- Both "master" and "use" images should be created under some circumstances. The master image will be of the highest quality and only used for creating new use images. Keep a master copy if any of the following are true:
    - The use images will be in a format not suitable for long-term preservation.
    - Different formats of use images need to be created (e.g., a GIF image for each page for on-screen display, and a PDF version for printing). Creating copies from a higher quality image will produce the best copies.
    - The printed original is being destroyed or is difficult to access.
- Software and hardware obsolescence - Some aspect of the formatting (most likely operating system or file type) or hardware (most likely disk type) will require migration within 5 years or so. Migration will add to the cost so should be factored into cost estimates.
- Certain media storage formats, like magnetic tape and removable disks (e.g. DVD) make it difficult to apply retention periods since the whole disk or tape has to be disposed of at once, even if individual records have different retention periods.
- Make backup copies
    - For removable disks, make at least 2 copies of each disk and keep them in separate, secure, locations. Removable disks, either magnetic or optical, can, and do, go bad with no notice, so keeping one copy is foolhardy, especially considering the expense of scanning.
    - Server-based records should be regularly backed up. Even then, though, a long-term backup is advisable since damage early in the storage period would produce many future damaged backups.
    - Check originals and backups regularly so that errors are discovered quickly.
- There's no easy answer for deciding when to migrate the records but it is simple: do it while you still can. A good rule of thumb is to review stored electronic records whenever the office changes its regular software.
- Storing long-term records only on removable disks (like CDs or DVDs) is generally not a good idea. Storing at least one copy online (e.g., a file server or web server) is recommended.
    - Long-term records will periodically need to be migrated to new formats and media. Migrating files kept on several removable disks is a tedious process.
    - Multiple copies of each disk must be created since removable disks can become unusable in a short period of time.
- Technical specifications for the master image will depend on the project, but here are some common specifications:
    - TIFF (Tagged Image File Format) is a broadly adopted file format standard applicable to black-and-white (1-bit), grayscale (8-bit), and color (24-bit) digital images.
        - for color images, 24-bit RGB without compression
        - for non-color images containing illustrations, 8-bit grayscale without compression
        - for non-color images containing only text and/or line art, 1-bit with ITU-6 (aka "Group 4") lossless compression
    - Capture at a high-enough dpi (dots per inch) level to render the image clearly but not higher than necessary since that will increase the cost of storage.
    - For further information, contact Records Management Services.

What are the security considerations?

For records that may be subject to audit or legal action, the process of preparing, scanning, and storing the records must be laid out firmly with proper security precautions applied. Process is everything in making electronic records hold up in court. Security precautions should include:

- personnel access restrictions
- proper metadata

- good scanning procedures
- quality control
- making sure that the images can't be tampered with after scanning.
- Consult *ANSI/AIIM TR 31-2004 - Legal Acceptance of Records Produced by Information Technology Systems* for information on the procedures necessary for creating and maintaining legally acceptable imaged records.

The drive within a scanner may retain copies of your images after the job is done. These drives should be wiped clean when the process is complete.

For certain types of information, permission from your CIO for a scanning project may be required.

For further information on security, check the Harvard Enterprise Security, especially the section on working with vendors.