OVERVIEW OF CURRENT DIGITIZATION PRACTICES

A WHITE PAPER FOR THE ENVIRONMENTAL PROTECTION AGENCY

Amy R. Berman

Donald F. Egan

Alan S. Linden

LMI

GOVERNMENT CONSULTING

JULY 2007

LMI

Overview of Current Digitization Practices[1]

JULY 2007

INTRODUCTION

This paper reviews digitization practices in use today, and emphasizes factors that may be applicable to library digitization efforts at the Environmental Protection Agency (EPA). The Stratus team identified these practices through market research of digital library practices used by various government agencies and universities and through reviews of our collective experience with comparable efforts in both the public and private sectors. This paper includes recommendations regarding technique, cost, and quality of work for digitization of hard-copy library materials, taking into account the various document types that may be digitized, including printed text, rare or damaged text, book illustrations, manuscripts, maps and other oversized items, photographic prints, transparencies, and microfilm. The paper addresses the following topics:

- Document preparation—the scanning process
- Processing requirements for scanning, including the types of scanners and scanning capture software
- Indexing, including assigning and capturing bibliographic data such as subject terms for cataloging and Online Computer Library Center (OCLC) services
- Storage and archiving
- Imaging and library standards and policies, including those established by the American National Standards Institute (ANSI), Association for Information and Image Management (AIIM), American Library Association (ALA), and Library of Congress
- Industry performance metrics
- Document enhancement software
- Staffing and training requirements.

[1] Notice: The views, opinions, and findings contained in this report are those of LMI and should not be construed as an official agency position, policy, or decision, unless so designated by other official documentation.

DOCUMENT PREPARATION

Taking time to properly prepare documents for scanning saves time and money in the digitization process. When preparing any type of paper document, such as reports or manuscripts for scanning, the following manual steps are typically taken:

- Remove the document from its binder or folder, or separate it from its binding (to avoid removing the binding from a valuable document, a book scanner can be used)
- Flatten dog-eared pages and ensure the proper orientation of all pages
- Remove paperclips, staples, sticky notes, and other items attached to the document
- Insert separator sheets with bar codes between document sections for indexing purposes.
- For document collections containing multiple sizes, grouping by size can lessen time needed to recalibrate scanners and cameras.

Standard: ANSI/AIIM TR15-1997, *Planning Considerations Addressing Preparation of Documents for Image Capture Systems*, should be followed when preparing documents for scanning.

PROCESSING REQUIREMENTS

Processing requirements address the choices of software and hardware, such as scanners, computers and monitors, resolution, speed, single- or double-sided scan, document delimiters, level of color accuracy, image quality, corrections criteria and procedures, type of character recognition, format, indexing, and standards regarding metadata.

Standards: ANSI/AIIM TR19-1993, *Electronic Imaging Display Devices* should be followed when selecting imaging devices and ANSI/AIIM TR34-1996, *Sampling Procedures for Inspection by Attributes of Images in Electronic Image Management (EIM)* should be followed when for sampling rules and quality assurance sampling rules on image quality control.

Scanners

In any digital imaging lab it is essential that the correct scanner be chosen to meet imaging needs. Several types of scanners—such as flatbed, drum, and film scanners—are available. High speed scanners are used for standard paper sizes typically batching documents to optimize scanner use. Libraries that have digitized library collection materials have found that there is no single scanner that solves all digitization needs and that a variety of scanners are needed to complete digitization tasks spanning multiple formats. The type and condition of a document drives the scanner selection. For example, a flatbed scanner may be used for fragile, unbound documents or for rescanning, but for non-fragile, unbound items, a high-speed duplex scanner is optimal. Specialty scanners for bound books are available though usually at a greater cost.

Scanner selection is based on a number of criteria, including the following:

- Volume (average number of pages and images to be scanned)
- Scanner duty cycle (average number of scans recommended for a scanner model)
- Need for color, black and white, or gray scale scans
- Resolution and format
- Document size
- Single or double sided (also referred to as simplex or duplex)
- Scanner warranty
- Maintenance requirements.

Table 1 summarizes recommended resolutions and bit depths for various document types. Resolution is measured by dots per inch (dpi). The information in the table is based on studies published by a few university libraries and government organizations.[2] The following subsections address three key scanner features in more detail.

Table 1. Recommended Imaging Requirements

Document type	Resolution	Bit depth
Books (text pages)	400 or 600 dpi (access quality) 600 dpi (preservation quality)	1 bit (black and white bitonal) 24 bit (color)
Rare/damaged printed text	300–600 dpi	8 bit (gray scale) 24 bit (color)

[2] Digital Library Federation, *Benchmark for Faithful Digital Reproductions of Monographs and Serials*, December 2002, Government Printing Office (GPO) *Specifications and Metrics for Quality Control of Converted Content,* March 2006, University of Virginia Library, *University of Virginia Community Digitization Guidelines*, March 6, 2006, and Western States Digital Standards Group, Digital Imaging Working Group, *Western States Digital Imaging Best Practices*, Version 1.0, January 2003.

Book illustrations or figures	400 dpi (access quality) 600 dpi (preservation quality)	8 bit (gray scale) 24 bit (color)
	300 (larger pages) or 400 dpi 300 dpi (with text)	8 bit (gray scale) 24 bit (color)
Manuscripts	300–600 dpi	8 bit (gray scale) 24 bit (color, if color present in original)
Maps and other oversized items	400 dpi	8 bit (gray scale) 24 bit (color)

Standard: ANSI/AIIM MS44-1998, *Recommended Practice for Quality Control of Image Scanners*, should be followed to ensure scanner quality control and continued maintenance of an established level of quality.

Color, Black and White, and Gray-Scale Scanners

Most scanners offer color features because the cost of color scanning has been radically reduced. In addition, storage costs per gigabytes (GB) have declined rapidly so storage capacity is less of a cost issue than in the past. Color gives a truer rendition of the document if colors are present in the original. Gray scale can be used to improve the scanned quality of low-contrast images.

Resolution

Resolution is the "density of pixels captured in the digitization of an image" when digitized.[3] Images of library materials can be captured at anywhere from 300 dpi to 600 dpi, depending on the nature of the documents. The resolution should be determined according to the type of document being scanned, with quality of the image taking precedence. The Library of Congress's standard is 300 dpi.[4] This resolution is also recommended by AIIM and should be considered as the minimum resolution for EPA scanning.

Bit depth refers to the number of colors that can be displayed. The higher the bit depth, the more color variation that can be captured and displayed.

Format

Documents are typically stored as Tagged Image File Format (TIFF), Portable Document Format (PDF), Portable Document Format for archiving (PDF/A), or Joint Photographic Experts Group (JPEG) files. The preferred format depends on

[3] University of Virginia Library, *Internal Production Digitization Standards*, March 6, 2006.

[4] Fleischhauer, Carl. *Digital Formats for Content Reproductions.* The Library of Congress. July 13, 1998.

the document type. Both TIFF and PDF/A formats will store a true facsimile image at no additional cost. Table 2 summarizes the advantages of TIFF and PDF/A formats.

Table 2. Advantages of TIFF vs. PDF/A

TIFF	PDF/A
Uses well-proven storage method for images and has been around longer Supports most current paper processing operations	Provides a stripped-down version of PDF, which is a proprietary format of the Adobe Corp. Is promulgated by various federal government agencies and is not proprietary, so would have no additional cost Is endorsed by significant standards bodies Provides the ability to store metadata along with scanned images Has been adopted by the U.S. Courts Contains its own description

The software is free and neither format has a cost advantage. Most of our research findings indicated libraries using the TIFF format for digitized library materials, but in many cases, those recommendations were published before the introduction and increased usage of PDF/A. Either TIFF or PDF/A format are acceptable formats to use for library materials.

JPEG is another file format that is used and may be considered for storing library digitized materials that are in color.

Other imaging formats (for example, GIF, JPEG and bitmap) are available, but these are not prevalent in the marketplace at this time.

Better Federal guidance on formats may emerge as the National Archives Records Administration has an active contract with Lockheed Martin to determine government standards for an electronic records archive. In the interim, we recommend following Library Congress practices and ISO 19005-1 standards.

Computers

The computer used for any scanning station must be able to handle very large files, which can be memory and processor intensive. Therefore, the computer should have adequate Random Access Memory (RAM) and disk space. The image-processing speed will have a direct impact on the workflow and the speed of the scans. A personal computer (PC) with a minimum of 2 GB of RAM should be used for image capture and quality control; this will also accommodate new operating systems software, which requires more RAM.

Monitors

Monitors are used to preview the quality of images to be captured. A large-screen monitor that supports a high-resolution display should be used for image editing and quality control. To properly review images, scanner users should have high-resolution screens of a minimum of 19 inches (flat panel preferred). The ability to calibrate and control the monitor's contrast, brightness, and color temperature is also important.

INDEXING

Indexing by record can be the most costly part of any digitization effort, yet is vital for allowing users to find the information they require.[5] Indexing can include assigning and capturing various metadata, including bibliographic data and subject terms. Deciding how library materials will be indexed prior to scanning is critical for future access to the digital images. This can possibly be automated by using bar codes and new Optical Character Recognition (OCR) formatting tools now available from vendors. The following are some ways that a document can be indexed:

- *Key from index fields.* This is done by manual entry and may include such information as the document type, date, document name, and subject matter. Not only must the data be captured, but it must be done in a standardized manner. Fortunately, in a library application these issues should have been addressed.

- *Bar codes for auto-indexing.* This requires collecting form information on a bar code before scanning a batch of documents and allows for automatic population. Another use of bar codes is as a quality control tool for generating statistics on misfeeds or double feeds. This works for larger volumes by randomly inserting a sheet with a certain number in the barcode. Doing this regularly can "guarantee that processes are checked against a certain percentage, without the hassle of comparing the originals with the scanned image."[6]

- *Zone OCR.* This is another mode for automatic text indexing. OCR software is used after scanning to distinguish between images and text and to establish what letters are denoted in the light and dark areas.

The type of document influences the bibliographic data captured. For example, the record for a monograph does not have to capture the issue and volume (but may have "Edition" information that needs to be recorded), but issue and volume

[5] Viera, Al. *The Evolution of Digitization – Web-Based Genealogy Spurs Digital Preservation.* TODAY – The Journal of Work Process Improvement, March/April 2007.

[6] AIIM, "Scanning and Capture Technologies 2007…Process Integration and ROI Enhancement," March 2007.

information is needed to retrieve periodicals and serials successfully. Table 3 contains typical metadata captured in various digital libraries.

Table 3. Typical Metadata

Field	Description
Title	A name given to the resource.
Creator	An entity primarily responsible for making the resource. Examples of a Creator include a person, an organization, or a service.
Subject	The topic of the resource. Typically the subject will be represented using keywords, key phrases or classification codes. Recommended best practice is to use a controlled vocabulary.
Description	An account of the resource. Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
Publisher	An entity responsible for making the resource available. Examples of a Publisher include a person, an organization or a service.
Contributor	An entity responsible for making contributions to the resource. Examples of a Contributor include a person, an organization or a service.
Date	A point or period of time associated with an event in the lifecycle of the resource. Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 (W3CDTF).
Type	The nature or genre of the resource. Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary (DCMITYPE). To describe the file format, physical medium, or dimensions of the resource, use the Format element.
Format	The file format, physical medium, or dimensions of the resource. Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types (MIME).
Identifier	An unambiguous reference to the resource within a given context.
Source	A related resource from which the describe resource is derived. The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
Language	A language of the resource. Recommended best practice is to use a controlled vocabulary such as RFC 4646.
Relation	A related resource. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

Table 3. Typical Metadata

Field	Description
Coverage	The spatial or temporal of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinated. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names (TGN). Where appropriate, named places or time period can be sued in preference to numeric identifiers such as sets of coordinates or date ranges.
Rights	Information about rights held in and over the resource. Typically, right information includes a statement about various property rights associated with the resource, including intellectual property rights.

Source: National Information Standards Organization (NISO), *The Dublin Core Metadata Element Set*, May 22, 2007.

Standard: ANSI/AIIM MS55-1994, *Information and Image Management: Recommended Practice for the Identification and Indexing of Page Components (Zones) for Automated Processing in an Electronic Image Management (EIM) Environment*, should be followed for indexing, especially for zoned OCR quality control.

Online Computer Library Center Services

Discovery services for locating and accessing EPA documents can also be provided by using OCLC services.

OCLC services include access to bibliographic, abstract, and full-text information when and where needed. OCLC is a “nonprofit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world’s information and reducing information costs.”[7] More than 57,000 libraries in 112 countries worldwide use OCLC services to detect, attain, catalog, loan, and preserve library materials. WorldCat, the OCLC Online Union Catalog, is a global database that contains bibliographic and library ownership information maintained by member libraries. OCLC services are being used by EPA and are an effective tool to allow discovery of EPA library materials.[8]

[7] OCLC Online Computer Library Center, http://www.oclc.org/.

[8] Transition Materials from EPA Library Network Intranet home. *EPA Library Transition: Repository Procedures*. July 2006 (Revised January 2007).

STORAGE AND ARCHIVING

When establishing a digitization capability, an organization must consider how it will store and retrieve electronic images, especially because digital images take up a large amount of space on any computer or network. Because the files will be online, access needs to be from a direct access storage device or large hard-disk drives. Hierarchical Storage Management (HSM) uses a combination of online, near-line and off-line storage. This can include CD, DVD, or other optical media; discs can be added as the file grows over time.

Metadata also needs to be taken into account when preparing digital images for storage and archiving. The following are some of the various types of metadata:[9]

- *Descriptive metadata* is used in the discovery and identification of an object. Examples are Dublin Core, VRA, and MARC records (OCLC uses the MARC record format). In addition, descriptive metadata for digital objects applies to information on the full collection of files associated with the digital object and their relationships to one another.

- *Structural metadata* is used to display and navigate a particular object for a user and includes the information on the internal organization of that object (for example, a book may have a preface, a table of contents, chapters, and an index).

- *Administrative metadata* represents the management information for this object. Examples of management information are the date the object was created, its content file format (e.g. TIFF, PDF/A, JPEG), scanning resolutions used, and rights information.

There are new OCR techniques for extracting metadata from free-form text. Although the new OCR techniques have not been proven in a library setting, it is being used in the commercial sector.

In addition to the data structures used to convey metadata, a number of standards, guides, and other resources are available for determining the semantics and syntax of the metadata content. These are usually application or data-format specific and include resources such as the following:[10]

- AACR2 (Anglo-American Cataloging Rules)

- LCSH (Library of Congress Subject Headings)

- AAT (Art and Architecture Thesaurus)
  http://shiva.pub.getty.edu/aat_browser/

[9] University of Kansas, Digital Initiatives, *Recommended Standards and Best Practices for Digital Projects*, January 7, 2003.

[10] See Note 6.

- TGN (Getty Thesaurus of Geographic Names) http://shiva.pub.getty.edu/tgn_browser/).

As much metadata that can be captured is recommended particularly when ease of retrieval is important. When capturing metadata, national and local standards should be followed.

When planning for storage and archiving, ISO 19005-1, *Document Management–Electronic Document File Format for Long Term Preservation*, *s*hould be referenced.

IMAGING AND LIBRARY STANDARDS

ANSI/AIIM, ANSI American Society for Quality (ASQ) and ISO have developed imaging and library standards for document preparation, image processing, and so on, that should be considered when digitizing documents. We also recommend that other standards be used for guidance to improve document preparation and overall quality control during the scanning process. Table 4 lists the ANSI/AIIM and other standards that should be followed.

Table 4. Recommended Standards

Identifier	Title and description
ANSI/AIIM TR15-1997	*Planning Considerations, Addressing Preparation of Documents for Image Capture*—for document preparation
ANSI/AIIM MS44-1988 (R1993)	*Recommended Practice for Quality Control of Image Scanners*—for scanner quality control to ensure continued maintenance of an established level of quality
ANSI/AIIM TR19-1993	*Electronic Imaging Display Devices*—for selecting imaging devices
ANSI/AIIM MS55-1994	*Recommended Practice for the Identification and Indexing of Page Components (Zones) for Automated Processing in an Electronic Image Management (EIM) Environment*—for zoned OCR quality control
ANSI/AIIM TR34-1996	*Sampling Procedures for Inspection by Attributes of Images in Electronic Image Management (EIM) and Micrographics Systems or ANSI Z1.4 Systems*—for sampling rules and quality assurance sampling rules on image quality control
ANSI/ASQ Z1.4-2003	*Sampling Procedures and Tables for Inspection by Attributes* – for quantifying performance.
ANSI/ASQ Z1.9-2003	*Sampling Procedures and Tables for Inspection by Variables for Percent Nonconforming* – for quantifying performance.
ANSI/NISO Z39.85-2007	*The Dublin Core Metadata Element Set* for defining metadata elements for resource descriptions.
ISO 19005-1	*Document Management–Electronic Document File Format for Long Term Preservation*—for storage (PDF/A)

In addition to the ANSI/AIIM standards, ALA has recently articulated polices that may be relevant and followed by EPA when digitizing library materials.. The following are examples:

- ALA Policy 51, Federal Legislative Policy – This policy addresses the Federal Government's Role in Library and Information Services.

- ALA Policy 52.2.1, Preservation Policy – This policy is based on ALA's goal to ensure that every individual has access to information whenever needed and in a usable format.

- ALA Policy 55, Standards Policies[11]. – This policy addresses standards and guidelines adopted by ALA for the delivery of library services. ALA classifies standards as "policies which describe shared values and principles of performance for a library[12]" and guidelines as "procedures that will prove useful in meeting the standards[13]".

- ALA Draft Principles for Digitized Content (just passed at the 2007 Conference this week)—This policy document formalizes some general principles related to digitization.

Legal considerations regarding copyright of materials should also be taken into account when selecting items for scanning and defining the metadata.

INDUSTRY PERFORMANCE METRICS

Three industry performance metrics—notably, costs, speed, and accuracy—should be considered when establishing a digitization capability.

Costs

The various studies published over the years on the cost and effectiveness of digitization show a vast range of digitization and storage costs and are dependent on a range of factors, including software, hardware, imaging and indexing processes and workflows and quality control, as well as the type of library material. Various metrics were shared in several presentations given at the National Initiative for a Networked Cultural Heritage Symposium held in April 2003.

At this symposium, Donald Waters identified the following three costs as barriers that must be addressed to achieve cost savings in digitization:[14]

- *Technology and workflow costs.* Workflow has helped to improve digitization practices. In addition, technology used in digitization, such as capture and markup of text, and OCR have lowered costs.

[11] "ALA Governing and Strategic Documents". .American Library Association.

[12] See Note 11.

[13] See Note 11.

[14] Donald Waters, "The Economics of Digitizing Library and Other Cultural Materials" (presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

- *Intellectual property costs.* There are various initiatives to address copyright issues in order to avoid lawsuits. These initiatives "demonstrate that communities of users and publishers can find ways to create the trust and goodwill needed to overcome the costly barriers of copyright and create highly useful digitized collections of research and educational materials."[15]

- *Institutional costs and variables.* Minimal attention is given to "how to approach technology or intellectual property costs factors."[16]

Two other presentations at the symposium, one by Nancy Harm and the other by Dan Pence, addressed various factors that can affect average costs.[17,18] Those factors include staffing and staff experience, equipment, workspace, type of binding (bound or unbound), page size, scanning resolution, scanning bit depth, handling requirements (fragile/not fragile), and place of performance (on or off site).

Pence's study did note that one metric that is difficult to quantify is that of vendors taking responsibility for a large number of potential risk factors, such as the following:[19]

- Equipment is fully utilized over 3 years.

- Equipment failure is minimal.

- Software upgrades do not cause problems.

- Supply of material is not interrupted.

- Employees show up for work.

- Employees maintain productivity.

- Employees handle material carefully.

- Image quality meets or exceeds requirements, resulting in little or no rework.

These risk factors are only applicable to EPA if the scanning of library materials is an EPA managed operation. If EPA is outsourcing their scanning operations then the vendor has the responsibility for these risk factors.

[15] See Note 14. (Should this be 14 rather than 9?

[16] See Note 14.   Same as above?

[17] Nancy Harm, "Luna Imaging: A Manufacturing Model" ((presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

[18] Dan Pence, "Ten Ways to Spend $100,000 on Digitization" (presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

[19] See Note 13.

Another study[20] reported cost figures from the Library of Congress that provides a point of reference for various types of digitization:[21]

- Base-level digitization: $5.32 per page; $1,600 for a 300 page book
- Enhanced digitization with full SGML encoding to aid full-text searching and analysis: $8.25 per page; $2,500 for a 300 page book
- Photograph images: $18.51 per image; $40,730 total costs for 2,200 images from a photograph collection
- Archival color slides: $20.13 per image; $301,937 total costs for 15,000 archival color slides
- Periodical index: $0.14 per page image; $32,760 for creation of 234,000 bitonal page images (text only)

Table 5 reports a break down of costs of various steps in the digitization process reported by a number of institutions in an article by Steve Puglia from National Archives and Records Administration (NARA). That variation in costs is significant.

Table 5. Cost Ranges Reported for Various Digitization Processes

Digitization category	Digitizing	Metadata creation	Other	Overall costs
Overall projections (per image)	$0.25–$19.80	$0.75–$34.65	$0.45–$50.20	$1.85–$96.45
Adjusted projections (per image)	$0.25–$16.65	$0.75–$17.25	$0.45–$28.15	$1.85–$42.45
Mixed collections (per item)	$3.45–$16.50	$2.85–$17.25	$4.50–$21.55	$3.25–$40.50
Single items (per page)	$1.90–$8.00	$5.75–$12.85	$7.60–$28.15	$23.10–$35.80
Photographs (per photo)	$2.30–$16.65	$4.85–$6.45	$3.35–$24.65	$5.20–$42.45
Books/pamphlets	$2.10–$6.10	$1.50–$11.10	$1.35–$6.90	$4.60–$14.40
Re-keyed text (per page)	$2.55–$5.00	$2.35–$5.70	Limited Data	Limited Data
OCR (per page)	$0.25–$3.60	$0.75–$2.40	$0.40–$2.10	$1.85–$7.65

Source: Puglia, Steven, National Archives and Records Administration (NARA). *The Costs of Digital Imaging Projects.* RLG DigiNews. October 15, 1999, Volume 3, Number 5.

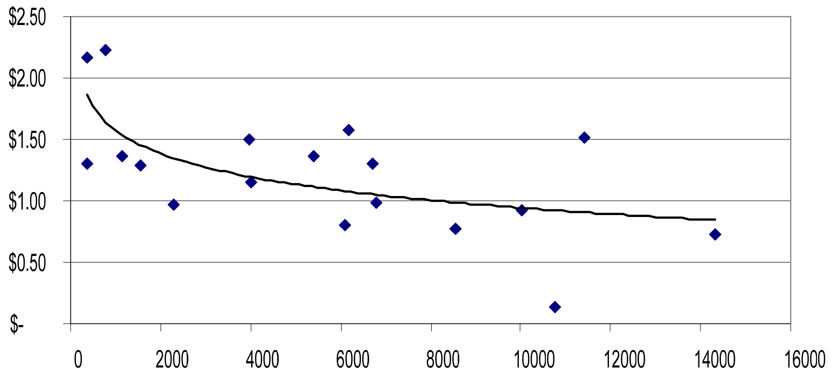
[20] "Digitization Costs and Funding" Digital Library Workshop (October 2, 2003).

[21] "The Evidence in Hand: Report of the Task Force on Artifact in Library Collections" *Library of Congress.* November 2001.

In addition, our research found that the cost of capture and conversion of data typically comprises one-third of the total costs. Cataloging, description, and indexing typically constitute two-thirds of the total costs. Cataloging costs may be optimized since OCLC is being used by EPA, although there will be some associated costs.

Figure 1 shows the digitizing costs per page for documents of different lengths. The costs were provided by various contractors as part of an LMI study for a government agency. The gap between the most-effective and the least-effective contractors, in terms of the processing costs, was significant. Costs are apparently independent of the input mechanism used: one high-cost contractor has been using imaging since 1994, but its per-document costs were $1.58, the second highest of the large contractors. However, our interview with that contractor indicated that it may be including mainframe correction processing costs that other contractors did not include in their costing submissions.

Figure 1. Cost per Paper Document



Speed

In a previous study performed by LMI, various contractors were interviewed who used scanners capable of processing 100 pages per minute or more for bitonal scanning. Some contractors were using scanners that were 5 or more years old, while others had fairly new devices. Those using newer scanners tended to have lower costs per page.

Accuracy

Contractors with the strongest quality control procedures initiated at the start of the scanning process are also those with the lowest costs. The contractors used various image enhancement tools, depending on the make and brand of the scanner and capture software. Quality control is essential in the overall scanning process and appropriate steps should be taken up front to ensure the right quality control process is in place.

DOCUMENT ENHANCEMENT SOFTWARE

Document enhancement is important for achieving labor savings and optimizing productivity. When document enhancement software products are used, less documentation preparation is required before scanning. As a result, labor hours are saved, since almost all rescans and operator interventions are eliminated. This is achieved by the various image processing capabilities such as white-level tracking, deskewing and despeckling, automatic color detection, threshold processing, and blank image detection and removal.

STAFFING AND TRAINING REQUIREMENTS

Staffing of the scanning facility with well-trained personnel is very important, since labor tends to be the largest of the ongoing costs for any document scanning implementation. An estimated 80 percent of the scanning cost is labor, so it is important to have any labor savings devices, such as faster scanners, OCR, use of bar codes, or anything else that makes the scanning process more efficient. In addition, it is advantageous to have personnel who have knowledge of cataloging, registration methods, and metadata schema. If outsourcing, selection of experienced, reliable, quality and efficient contractor support is key.

Sufficient time for training and opportunities to receive further education and training should also be provided, especially if there is any change in the scanning process or new technology or hardware.

Experienced end users understand that obtaining (and maintaining) committed employees is a challenge. This comes as no surprise because organizations traditionally spend a disproportionate share on the technology itself and tend to underspend on the training and education needed to maximize the full potential of these technologies.[22]

Training can be provided by a variety of organizations, including the following:

- AIIM
- ALA
- Library and Information Technology Association
- Association for Library Collections and Technical Services
- Society of American Archivists
- Institute for Museum and Library Services.

[22] AIIM, “Scanning and Capture Technologies 2007…Process Integration and ROI Enhancement,” March 2007.

SUMMARY

An imaging strategy that establishes standard imaging practices resulting in more efficient imaging is essential when transitioning to a digitized library. Imaging efficiency is dependent on technology, trained personnel, and experienced management. Decisions on scanning and indexing should be coordinated in advance to correlate with storage and retrieval strategies.

Over the last 5 years, scanning technology has improved dramatically, as has the technology for enhancing images and OCR processing. The technology for capture processing, including dedicated workflow structures for image processing, also has improved. By implementing current but proven technology, costs can be reduced substantially. Having a well-trained and dedicated workforce with a good operating environment is essential, as is strong management with experience in processing large volumes of paper images. In addition, we recommend EPA take into consideration the quality control process, indexing and resolution.

REFERENCES

AIIM Electronic Content Management Conference (Boston, MA, April 2007).

AIIM, "Scanning and Capture Technologies 2007" Process Integration and ROI Enhancement," March 2007.

ALA OITP Workshop on Mass Digitization, *Policy Areas for Revision: Digitization of Library Resources*, August 2006.

American Library Association, Office of Information Technology Policy Advisory Committee, Digitization Policy Workgroup, "Digitization Policy Areas" (results, Digitization Policy Workshop, Chicago, IL, April 2006).

Arms, Caroline R., "Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress", *D-Lib Magazine*, April 1996.

Biknese, Douglas A., *Measuring the Accuracy of the OCR in the Making of America* (University of Michigan, School of Information, Winter 1998).

Chapman, Stephen, "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?" *Journal of Digital Information*, Vol. 4, No. 2 (February 2003).

Chester, Bernard, "Archiving Electronics Files," *AIIM E-DOC Magazine*, May/June 2006.

Dempsey, Kathy, "A Dozen Primers on Important Information Standards," *Computers in Libraries*, April 2007.

Digital Library Federation, The Digital Library Federation Benchmark Working Group, *Benchmark for Faithful Digital Reproductions of Monographs and Serials*, December 2002.

Digital Library Federation, Council on Library and Information Resources, *Imaging Systems: The Range of Factors Affecting Image Quality*, 2000.

Digital Library Federation, Council on Library and Information Resources, *Measuring Quality of Digital Masters*, 2000.

Digital Library Federation, *Digital Library Standards and Practice*s (DLF Publications, Working Papers, Reports, and Other Digital Library Information Resources).

Duhon, Brian, "Capture—On-Ramp to ECM," Supplement to *AIIM E-DOC Magazine*, 2007.

Fleischhauer, Carl. *Digital Formats for Content Reproductions.* The Library of Congress. July 13, 1998.

Grogg, Jille E. and Beth Ashmore, “Google Book Search Libraries and Their Digital Copies,” *Searcher–The Magazine for Database Professionals*, April 2007.

Harm, Nancy, “Luna Imaging: A Manufacturing Model” ((presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

Harvard University Library, LDI Project Team, *Measuring Search Retrieval Accuracy of Uncorrected OCR: Findings from the Harvard-Radcliff Online Historical Reference Shelf Digitization Project*, August 2001.

Hughes, Lorna, “The Price of Digitization: New Cost Models for Cultural and Educational Institutions” (presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

International Federation of Library Associations, *Guidelines for Digitization Projects*, March 2002.

Jones, Ruth Ann, “Behind the Scenes in a Digitization Project,” http://www.library.wisc.edu/libraries/womensstudies/fc/fcjones222.htm, April 13, 2007.

Kructhen, Dolores, *The New Science of Document Imaging Delivers Results*. Document Imaging Eastman Kodak (brochure provided at AIIM Expo Today, April 19 2007).

Library of Congress, “Digital Collections and Programs: Library Functions,” http://www.loc.gov/library/library.digital.html, April 19, 2007.

Mancini, John, “AIIM Industry Watch Survey Scanning and Capture Technologies 2007” (AIIM, January 2007).

McClure, Marji, “Case Study: Rolling Out a Robust Library Experience Online,” *Information Today*, Vol. 24, No. 4 (April 2007).

MINITEX/LDS Joint Standards Review Task Force, “Guide to Digital Projects,” *Standards and Guidelines for Automated Library Systems*, Fall 2003/Winter 2004.

National Information Standards Organization*, A Framework of Guidance for Building Good Digital Collections*, 2nd Edition, 2004.

National Information Standards Organization (NISO), *The Dublin Core Metadata Element Set*, May 22, 2007.

Pence, Dan, “Ten Ways to Spend $100,000 on Digitization” (presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

Porter-Roth, Bud, “Document Backfile Conversion Project Guidelines,” *AIIM E-DOC Magazine*, January/February 2006.

Puglia, Steven, National Archives and Records Administration (NARA). *The Costs of Digital Imaging Projects.* RLG DigiNews. October 15, 1999, Volume 3, Number 5.

Shufeldt, Laurie, “Connecting the Dots: Integrate Scanning with Data,” *AIIM E-DOC Magazine*, March/April 2006.

The University of Tennessee Libraries, *Report to the Digital Library Federation*, Fall 2003.

U.S Government Printing Office, *GPO LOCKSS Pilot: Final Analysis*, April 12, 2007.

University of Illinois at Urbana-Champaign, University Library, Digital Imaging and Media Technology Initiative, *Guidelines for Digital Imaging Projects*, December 6, 2001.

University of Kansas, Digital Initiatives, *Recommended Standards and Best Practices for Digital Projects*, January 7, 2003.

University of Virginia Library, *Internal Production Digitization Standards*, March 6, 2006.

University of Virginia Library, *University of Virginia Community Digitization Guidelines*, March 6, 2006.

Viera, Al, “The Evolution of Digitization: Web-Based Genealogy Spurs Digital Preservation,” *TODAY, The Journal of Work Process Improvement*, March/April 2007.

Vise, David A., “World Digital Library Planned,” *The Washington Post*, November 22, 2005.

Washington State Library, “Digital Best Practices,” http:/digitalwa.statelib.wa.gov/best.htm.

Waters, Donald, “The Economics of Digitizing Library and Other Cultural Materials” (presentation, National Initiative for a Networked Cultural Heritage Symposium, April 8, 2003).

Western States Digital Standards Group, Digital Imaging Working Group, *Western States Digital Imaging Best Practices*, Version 1.0, January 2003.