FINAL

DIGITIZATION ASSESSMENT

A WHITE PAPER FOR THE ENVIRONMENTAL PROTECTION AGENCY

Amy R. Berman

Donald F. Egan

Alan S. Linden

.

LMI
GOVERNMENT CONSULTING

MARCH 2008

The views, opinions, and findings contained in this report
are those of the Stratus Team, consisting of LMI and Stratus Consulting, and should not be construed as
an official agency position, policy, or decision,
unless so designated by other official documentation.

LMI

Digitization Assessment

FEBRUARY 2008

BACKGROUND

The President's budget for the Environmental Protection Agency (EPA) proposed a $2 million cut in FY 2007, directed at the Agency's Regional and Headquarters library. In anticipation of this reduction, some libraries closed access to their physical space at the end of FY 2006 and the beginning of FY 2007. Libraries being considered for closure were instructed to box library materials that are not in the EPA National Environmental Publication Information System (NEPIS) database and to send them to a contractor in Cincinnati, OH, for digitization and entry into NEPIS.[1] During 2006, the contractor digitized 15,260 documents, consisting of 1,260,000 pages. A new contractor, located in Rockville, MD, took over the project at the beginning of 2007.

EPA asked Stratus Consulting to perform an independent assessment of the digitization process,[2] because questions about the process arose, especially during the period when the contract was recompeted and a new contractor selected. As the first step, EPA asked the Stratus team to perform market research and to document current digitization practices.[3] That work, which included a review by an independent third-party digitization expert and librarian from academia, is documented in *Overview of Current Digitization Practices* (July 2007). EPA then asked the Stratus team to compare the EPA and contractor plans to digitize the library documents against practices identified through market research, identify standards that the contractor should follow, and recommend improvement actions. This paper conveys the results of that task.

The findings in this paper are based on two interviews, one with staff members from EPA's Information Resources Management Division on June 19, 2007, and one with EPA's digitization contractor on August 31, 2007. This assessment is based purely on anticipated practices conveyed by EPA and its contractor, not on observation of current practices. This paper also draws heavily from *U.S. EPA Library Collections Digitization Process Report*, submitted to EPA by the digitization contractor in September 2007.

[1] EPA, *FY 2007 Library Plan: National Framework for the Headquarters and Regional Libraries*, August 15, 2006.

[2] Stratus Consulting subcontracted with LMI to perform this assessment due to LMI's subject matter expertise and experience. Stratus Consulting oversaw the task and reviewed deliverables.

[3] The Stratus team comprises LMI and Stratus Consulting.

At the time of these interviews and the publication of the digitization process report, only 15,260 EPA library documents had been scanned during the 06-07 timeframe, thus the majority of the information contained in this report is based on plans for the digitization of EPA library documents.

This paper begins with an overview of the digitization process and prescanning preparations. It then addresses the following areas:

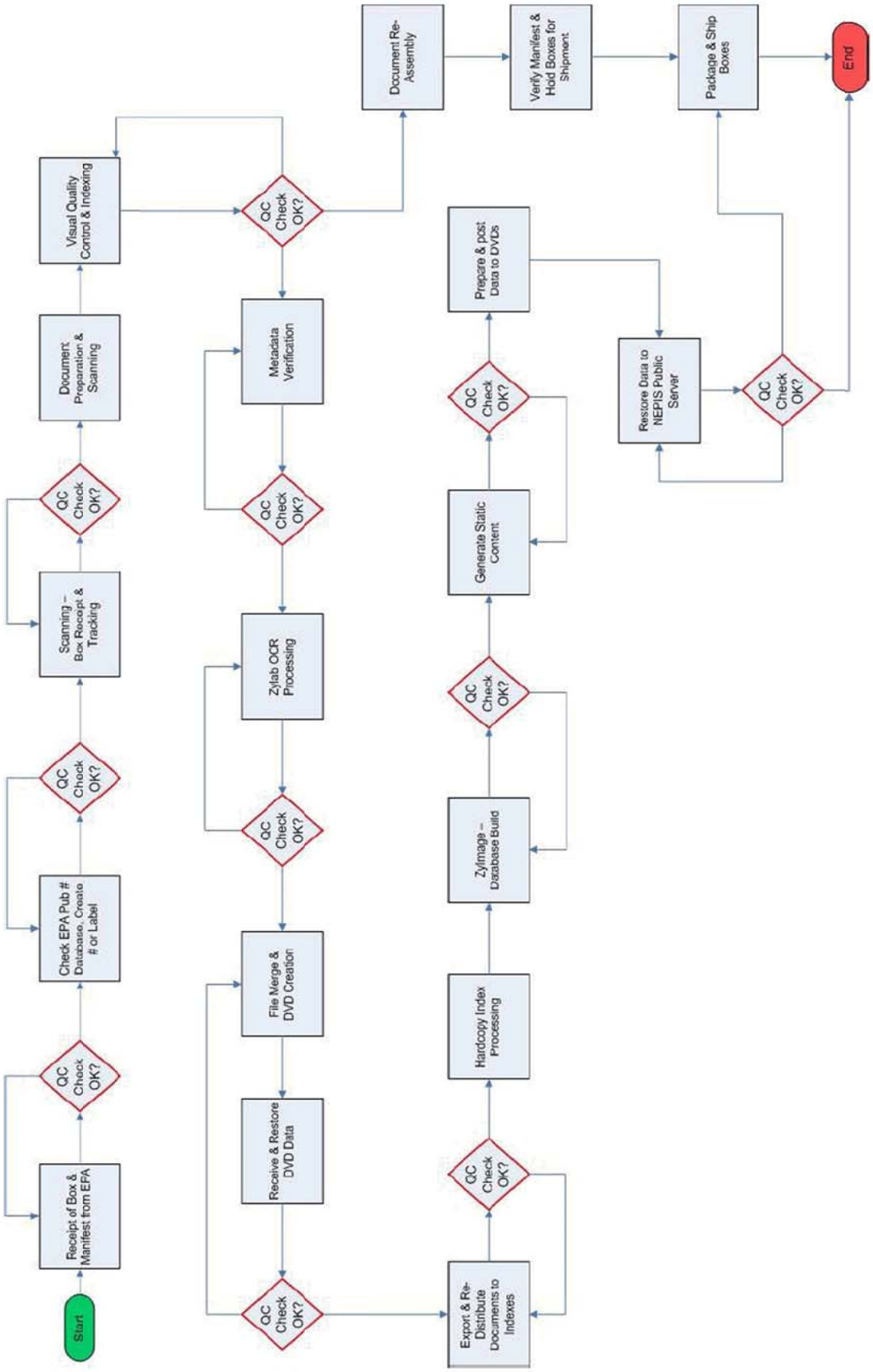
- Document preparation
- Hardware and software
- Indexing
- Storage and archiving
- Imaging and library standards and policies
- Industry performance metrics
- Quality control
- Document enhancement software
- Optical character recognition (OCR)
- Staffing and training.

DIGITIZATION PROCESS OVERVIEW AND PRESCANNING PREPARATIONS

Figure 1 shows the digitization workflow, from receipt of the library collection boxes to the loading of the images and data onto the EPA NEPIS website, which the contractor plans on following.[4] The following sections describe the various processes and steps planned for the digitization of EPA library materials.

[4] EPA, *U.S. EPA Library Collections Digitization Process Report*, September 24, 2007. [http://www.epa.gov/nscep/DigitizationReport.pdf]

Figure 1. Document Digitization and Management Process Flow



Source: EPA, *FY 2007 Library Plan: National Framework for the Headquarters and Regional Libraries*, August 15, 2006, p. 3, Exhibit 1.

Before sending documents to the contractor, EPA librarians will be responsible for printing out a copy of the On-line Library System (OLS) record for each document and placing it inside the front cover of each document. Librarians will also be responsible for packing documents in appropriate boxes for shipping, preparing a manifest of the documents in each box, and placing the manifest on top of the documents in the box. Each box will include the name of the originating library, a box number, and instructions indicating whether the documents are to be sent to a repository library or to the originating library after digitization. The boxes will then be sent to the contractor for scanning.[5]

Upon receipt of the boxes of library documents, the contractor will be responsible for logging each box into an EPA Library Document Log, tracking each box, and checking the contents of the boxes to ensure that they match the manifest. If discrepancies are found, the originating library will be contacted. Each document will be checked to verify that the assigned EPA publication number is "properly formatted and standardized."[6] The title on the physical document will be compared to the title on the manifest, and if discrepancies are found, the manifest will be updated.[7]

DOCUMENT PREPARATION

Prior to scanning, the contractor will prepare the documents in batches to facilitate high-speed scanning and maintain the physical integrity of batches in the document collections.[8] The contractor plans to develop instructions based upon EPA-specified requirements to ensure maintenance of the collation and integrity of the document batches, including the original source file configurations, throughout the document conversion process. Line supervisors will ensure that members of the document preparation staff follow the instructions contained in a Document Imaging Work Order.[9]

In those cases where EPA chooses to have documents requiring special handling scanned, the contractor does have procedures in place. Document preparation manuals have been developed, by the contracto,r that contain specific instructions for how they will manage documents that require special handling, such as "onionskin paper, brittle, fragile, damaged, old, or one-of-a-kind documents. Amongst other steps taken for special handling if required, the contractor's personnel will wear latex gloves when handling and processing these types of documents."[10]

---

[5] EPA, *EPA Library Transition: Repository Procedures*, Attachment 1, Procedures for Digitizing and Dispersing Library Collections, July 2006 (updated January 26, 2007).

[6] Note 4, p. 2.

[7] Note 4, p. 5.

[8] Note 4, p. 5.

[9] Note 4, p. 5.

[10] Note 4, p. 5.

During the initial batch of EPA library documents scanned in January 2007, only about 1 percent of the documents were fragile. The initial batch contained a variety of collection presentations in terms of documents with clips or staples and different types of documents, but all were in generally good condition.

The contractor plans to follow the document preparation process recommended by the American National Standards Institute (ANSI) and Association for Information and Image Management (AIIM). The recommended process, articulated in ANSI/AIIM standard TR15-1997, includes the following actions:

- Remove any binding elements, such as staples, paper clips, prong fasteners, metal slide clasps, and any other items attached to the document[11]
- Insert separator sheets with bar codes between document sections for indexing purposes
- Use a paper cutter to remove the spine from bound publications in preparation for scanning
- Flatten dog-eared pages and ensure the proper orientation of all pages.

Recommendation: For the normal course of business, latex gloves are appropriate, but cloth gloves are recommended for handling archival or brittle materials.

Standards to Follow: During its document preparation activities, the contractor will follow, where applicable, ANSI/AIIM TR15-1997, "Planning Considerations Addressing Preparation of Documents for Image Capture Systems," and ANSI/AIIM MS52-1991, "Recommended Practice for the Requirements and Characteristics of Original Documents Intended for Optical Scanning."

HARDWARE AND SOFTWARE

This section addresses the hardware and software that the contractor will be using for scanners, computers, and monitors. It also includes findings regarding specific parameters that will be used, including 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 metadata standards.

Standard to Follow: The contractor will follow ANSI/AIIM TR19-1993, "Electronic Imaging Display Devices," when selecting imaging devices.

[11] Note 4, p. 5.

Document Scanners

This section addresses the types of scanners that will be used and the resolution, speed, and level of color accuracy expected.

The contractor plans to use a variety of scanners to support different types of documents and support high-volume operations, including high-speed automatic feed scanners and flatbed scanners for fragile, unbound documents or for rescanning. The contractor owns these scanners, having purchased them to fulfill the needs of other customers with similar requirements. The contractor will use appropriate scanners in accordance with the materials they are handling, as prescribed in their procedures.

Although the scanners have the capability of 400 to 800 dots per inch (dpi), documents will be scanned at 300 dpi. This resolution is specified by EPA and recommended by AIIM. The scanners being used by the contractor can scan according to customer needs in bi-tonal (black-and-white), grey-scale, or color. The scanners can also accommodate single- or double-sided scanning.

At this point, no daily scanning capacity has been established, but the contractor has the capacity and numbers of scanners required to meet scan demands consistent with EPA requirements. No capacity standards have been established under the EPA contract.

Table 1 lists the scanners that will be available for scanning EPA library documents and the potential resolution and pages per minute.

Table 1. Scanners Available for Scanning

Model	Rated speed (pages/minute)	Resolution[a]
Kodak i600 (black-and-white and color)	120	Up to 800 dpi
Fujitsu fi-4750C (black-and-white and color)–has flatbed capability	90	Up to 800 dpi
Fujitsu M4097D–has flatbed capability	90	Up to 600 dpi
Kodak 1500	50	Up to 600 dpi
Kodak 3520	90	Up to 600 dpi

[a] Rated speed slows down as resolution increases.

Standards to Follow: The contractor will follow ANSI/AIIM MS44-1998, "Recommended Practice for Quality Control of Image Scanners," to ensure scanner quality control and continued maintenance of an established level of quality; ANSI/AIIM MS53-1993, "Recommend Practice; File Format for Storage and Exchange of Image; Bi-Level Image File Format: Part 1," and ANSI/AIIM TR19-1993, "Electronic Imaging Display Devices," for selecting imaging devices; and ANSI/AIIM TR26-1993, "Resolution As It Relates to Photographic and Elec-

tronic Imaging," for evaluating photographic and electronic imaging products. In addition, we recommend consulting the NARA and Digital Library Federation websites for guidelines and best practices regarding document scanning.[12]

Format

Documents will be stored as Tagged Image File Format (TIFF) files. This format was specified by EPA and is an established industry standard for digitizing library materials.

EPA has considered storing documents in PDF/A format, but has not implemented it because it is not yet an approved standard at EPA. This may change as better federal guidance on formats emerge. For example, the National Archives and Records Administration (NARA) has a contract to determine government standards for an electronic records archive. NARA is also part of the PDF/A standards committee.

Table 2 compares the advantages of using the TIFF versus the PDF/A storage format.

Table 2. Comparison of TIFF and PDF/A Storage Formats

TIFF	PDF/A
Uses well-proven storage method for images and has a long service record	Provides a less complex version of PDF, which is a proprietary format of the Adobe Corp.
Supports most current paper processing operations	Is promulgated by various federal government agencies and is not proprietary, so would have no additional cost
Provides the ability to store metadata along with scanned images	Is endorsed by significant standards bodies, such as ISO
Accepted as a long term preservation standard in many library/archive environments	Provides the ability to store metadata along with scanned images
	Has been adopted by the U.S. Courts
	Contains its own description

Recommendation: EPA should give serious consideration to adopting PDF/A as its archival standard for library documents if it is approved by NARA.

Computers and Monitors

The contractor will be using Dell PC-Optiplex GX270 computers for its scanning stations with 1 GB of RAM. In our initial white paper, the Stratus team recommended that computers with a minimum of 2 GB of RAM be used for image capture and quality control.

[12] The NARA website can be accessed at http://www.archives.gov/, and the Digital Library Federation website can be accessed at http://diglib.org/standards/bmarkfin-draft.htm.

The contractor will be using dual 19-inch monitors to preview the quality of images to be captured and review them for completeness. This adheres to the Stratus team's recommendation and will allow the contractor to have a large enough monitor to properly review images.

Recommendation: The contractor could consider upgrading to 2 GB of RAM to accelerate image processing. As the 19-inch monitors age, we recommend that over time, the monitors be upgraded to 21-inch flat-panel high-resolution monitors,which are more readable, have a smaller footprint, and use less energy than CRT-based monitors.

Software

The proprietary scanning software being used by the contractor has on-screen displays that will enable the operator to make the image-setting adjustments necessary to optimize the scanner's output.[13] The scanner operator should be able to identify and rectify quality issues such as misfed pages, poor image contrast, and incomplete images.[14] If the scanner operator catches errors during this step, he or she can adjust the scanner's settings to respond to the changing document conditions and sizes, and then rescan the documents immediately.[15]

INDEXING

The contractor is using the Zylab suite of software products for capture, OCR, indexing, search, and retrieval. This includes the use of the Zylab product ZySCAN. ZySCAN allows for both automatic and manual indexing. ZySCAN recognizes all text in a document, regardless of the size, and automatically indexes the information in the document based on the recognized text.[16]

After image inspection, the contractor will index the metadata from the image collection. The data elements specified in the project requirements will be captured using separator sheets during the document preparation phase. "Document metadata, such as the EPA publication number, tile, number of pages and date created earlier in index files will be presented in the indexing module for verification and correction. The original manifest data is used for validation."[17]

The contractor will index fields specified by EPA. Currently, the only required fields (metadata) are the EPA publication number, EPA publication title, number of pages in each document, source (hard copy, digital, both), document type, scanning operator, and the scan date and publication year (in a four-digit format).

[13] Note 4, p. 6.
[14] Note 4, p. 6.
[15] Note 2, p. 6.
[16] Zylab Product Sheet, www.ZYLAB.com.
[17] Note 2, p. 8.

Only the EPA publication number, publication title, and publication year have been entered into the Zylab software and database.

Recommendations: An EPA goal for FY08 is to work with the OLS to determine the best approach for metadata creation. The contractor has the capability and flexibility to index as needed, so can accommodate any metadata changes. We concur with EPA's plan to improve the metadata structure and strongly encourage EPA to make this a high priority.

EPA should also consider including the metadata typically captured in digital libraries. Table 3 lists and describes the fields.

Table 3. Typical Metadata

Field	Description
Creator	An entity primarily responsible for making the resource. Examples of a creator are 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 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 are a person, an organization, or a service.
Contributor	An entity responsible for making contributions to the resource. Examples of a contributor are a person, an organization, or a service.
Format	The file format, physical medium, or dimensions of the resource. Examples of dimensions are 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 described 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.
Coverage	The spatial or temporal topic 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 coordinate. 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*. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.

Table 3. Typical Metadata

Field	Description
Rights	Information about rights held in and over the resource. Typically, rights information includes a statement about various property rights associated with the resource such as intellectual property rights.

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

Standard to Follow: For indexing, especially for zoned OCR quality control, the contractor will follow 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."

STORAGE AND ARCHIVING

Images are stored and can be retrieved by EPA employees and the general public via the NEPIS/National Service Center for Environmental Publications public website. This site allows access to EPA products free of change.

We discussed the level of capability of NEPIS to protect the electronic documents. EPA personnel believe that with the EPA's disk-to-disk to tape backup process and dual-site environment, NEPIS has sufficient capability to protect against loss of data. Furthermore, EPA believes that its selection is a sustainable data format and that the recent hardware refresh (and refreshes that will occur over the lifetime of NEPIS) will minimize the likelihood of magnetic drift in the storage media.[18]

Recommendations: To better protect its data and improve its sustainability, EPA should consider using PDF/A, if it is endorsed by NARA, and Hierarchical Storage Management (HSM).We suggest these as overall archival procedure. HSM uses a combination of online, near-line, and off-line storage and can save money. This can include CD, DVD, or other optical media; disks can be added as the file grows.

Standards to Follow: The contractor will follow ISO 19005-1, "Document Management–Electronic Document File Format for Long Term Preservation," for storage and archiving.

[18] Interview with personnel from EPA's Information Resources Management Division on June 19, 2007.

IMAGING AND LIBRARY STANDARDS AND POLICIES

The EPA contractor is adhering to many of the imaging and library standards developed by ANSI/AIIM and ISO for document preparation, image processing, and other activities that were recommended in *Overview of Current Digitization Practices*. Table 4 lists the ANSI/AIIM and other standards that are being followed by the EPA contractor.

Table 4. Standards EPA Contractor Adheres to

Identifier	Title and description
ANSI/AIIM TR15-1997	*Planning Considerations, Addressing Preparation of Documents for Image Capture*—for document preparation
ANSI/AIIM TR19-1993	*Electronic Imaging Display Devices*—for selecting imaging devices
ANSI/AIIM TR26-1993	*Resolution As It Relates to Photographic and Electronic Imaging*—for evaluating photographic and electronic imaging products
ANSI/AIIM TR31-2-1993	*Performance Guideline for the Legal Acceptance of Records Produced by Information Technology Systems, Part 2: Acceptance by Government Agencies*
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/AIIM TR38-1996	*Compilation of Test Targets for Document Imaging System*
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 MS52-1991	*Recommended Practice for the Requirements and Characteristics of Original Document Intended for Optical Scanning*
ANSI/AIIM MS53-1993	*Recommend Practice; File Format for Storage and Exchange of Image; Bi-Level Image File Format: Part 1*
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

Recommendation: We recommend that EPA consider having the contractor adhere to additional standards to further improve document preparation and overall quality control during the scanning process. In addition, EPA should have the contractor consult the NARA and Digital Library Federation websites for guidelines and best practices regarding digitization practices. Table 5 lists our recommended standards.

Table 5. Standards Contractor Should Consider

Identifier	Title and description
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)

Note: ASQ = American Society for Quality, and NISO = National Information Standards Organization.

INDUSTRY PERFORMANCE METRICS

EPA did not establish any metrics that it expects the contractor to adhere to for the current digitization process and does not require the contractor to collect or report metrics. However, the EPA contractor estimates typical OCR scanning accuracy rates at 99 percent, based on experience with similar customers.

Recommendation: The Stratus team recommends that EPA establish cost, speed, and volume metrics and require the contractor to report at least monthly once scanning operations resume.

QUALITY CONTROL

Quality control is an important part in a digitization effort. Figure 1 shows the various quality control steps taken throughout the EPA contractor's digitization process. The following is an overview of these steps, which the contractor has also used for other customers:

- *Document receipt.* The contractor will ensure that the documents have arrived as intended and that they have the information required for indexing.
- *Before scanning.* The documents will be checked to ensure that they are ready for scanning. This involves checking each document to make sure a "properly formatted and standardized EPA publication number" was assigned.[19]
- *Supervisory check.* The supervisor will perform a 100 percent review behind the line operator to ensure that each document is complete and ready.

[19] Note 4, p. 2.

- *Initial scanning.* Every document's image will be displayed on the monitor when scanned and reviewed by the operator. Operators may fix, rescan, insert, and delete pages.

- *After scanning.* Quality control will be conducted during which the original document will be compared with the scanned image. This involves matching the image with the original document.

- *Deliverable.* When images are burned to CD /DVD or other media, such as external hard drives, the contractor will run an electronic image check for compression and decompression errors and write errors.

The EPA contractor plans to develop a quality control plan that identifies quality assurance functions, quality control processes and measures, and overall quality assurance program responsibilities for each program.[20]

Recommendation: The steps planned should ensure quality control throughout the process. However, the contractor has not established a statistical approach for the random sampling to check for quality. While we are not aware of the terms of the EPA contract the Stratus team recommends implementing the statistical standards of the Institute of Electrical and Electronics Engineers or ANSI/AIIM for handling the quality checking of the OCR process which is used in many scanning operations. Currently, the contractor plans for 100 percent quality control during the OCR process, but to speed up the process, the Stratus team recommends that EPA consider using a random sample according to either AIIM or ANSI sampling procedures proposed in Tables 4 and 5. Random sampling is considered a best practice, since it is much more cost effective for checking OCR, because the cost of correction if checking 100 percent tends to be very expensive.

DOCUMENT ENHANCEMENT SOFTWARE

The contractor will use proprietary document enhancement software, which should help achieve labor savings and optimize productivity. As mentioned in *Overview of Current Digitization Practices*, when document enhancement software products are used, less documentation preparation is required prior to scanning. This is due to very minimal rescans or operation interventions being required, so labor hours are saved.

The contractor will process all scanned images through an image optimizer program that will be run on a dedicated image-enhancement server. Imaging technicians will use the various functions of the program, such as black border cropping, deskewing, rotation, despeckling, automatic character repair, and line removal to improve overall image quality. Image enhancement will provide clearer images for viewing and a better source image for improved OCR accuracy. Image enhancement procedures will not change the informational content of

[20] Note 4, p. 20.

the documents. To ensure legal admissibility, the EPA contractor plans to use control logs to record the file names of all images that are enhanced.[21]

Standards to Follow: The EPA contractor will follow ANSI/AIIM TR31-1993, "Performance Guidelines for the Legal Acceptance of Records Produced by Information Technology Systems."

OPTICAL CHARACTER RECOGNITION

The EPA contractor will use an advanced custom-developed OCR product, based on a commercial off-the-shelf product, to ensure the production and delivery of optimum-quality text files with the best interpretation of the original source material.[22] All documents will go through OCR.

Previously scanned document digital image data will be transformed into computer-readable ASCII text using Zylab OCR software. Rather than a single-engine OCR, Zylab uses multiple OCR engines with voting procedures to enhance accuracy. It identifies and coordinates each character on the image to capture the OCR. The OCR volume version of a deliverable image volume has a one-for-one image-to-OCR-text correlation in a matching directory structure. This process will ensure the order, integrity, and accountability for all source material while delivering a one-for-one image-to-text product.[23]

OCR accuracy is better than 95 percent. However, to compensate for those errors, EPA, when searching the NEPIS database, should consider the use of fuzzy logic or conceptual search tools for its public access system.

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

Recommendations: For improved searching, a conceptual search software should be used to compensate for OCR errors. New OCR software (beyond zonal definition) allows for the extraction of metadata from full text. EPA should consider using this software, because it will increase processing and throughput performance.

Standards to Follow: The EPA contractor will follow ANSI/AIIM TR31-1993, "Performance Guidelines for the Legal Acceptance of Records Produced by Information Technology Systems."

For indexing, especially for zoned OCR quality control, the contractor will follow ANSI/AIIM MS55-1994, "Information and Image Management: Recommended

[21] Note 4, p. 8.

[22] Note 4, p. 9.

[23] Note 4, p. 10.

Practice for the Identification and Indexing of Page Components (Zones) for Automated Processing in an Electronic Image Management (EIM) Environment."

STAFFING AND TRAINING

The contractor will maintain the following staff roles to manage the digitization process:[24]

- Project manager—provides overall project management support.
- Program quality assurance manager—provides overall program quality assurance.
- Program team—conducts up-front document review.
- Documentation preparation personnel—prepare documents for scanning.
- Line supervisors—supervise each part of the process.
- Scanner operators—conduct initial image quality review.
- Scanning supervisor—supervises the scanning operator and checks for work quality of the scanner operator by viewing random individual image files remotely.[25]
- Quality control operator—checks overall quality (including image quality and image size), ensures that all pages were scanned in sequence, and performs corrective rescan actions.[26] This also includes using the original manifest data for validation. "The index code representing each data element pertaining to a given image range is verified and flagged with the code. By comparing information available from the image of the document, the document indexers confirm the accuracy of the metadata."[27]
- Quality control supervisor—supervises the quality control operator and ensures that a final quality control check is conducted.
- Imaging technicians—conduct post-scanning processing of the images.

[24] Information provided during the interview with the EPA contractor and EPA staff members on August 31, 2007, and the *U.S. EPA Library Collections Digitization Process Report*, September 24, 2007.

[25] Note 4, p. 7.

[26] Note 4, p. 8.

[27] Note 4, p. 9.

- Distribution personnel——are responsible for written procedures and oversee the buddy system (the contactor's on-the-job training technique).

- Shipping specialists—return the original library materials after scanning.

When a new employee is hired by the contractor, the employee generally begins working in document preparation, because this area requires less skilled support. After gaining experience in document preparation, staff members are trained on the imaging process. All personnel are trained using user manuals, peer input, buddy system, days/hours of instructor-led training, and ongoing review. Over time, staff members are comprehensively cross-trained on the processes. Any individual not performing adequately is retrained. Some staff members are trained to specialize in quality control and distribution.

Every part of the digitization process will have a line supervisor, with no more than six people per group. On average, line employees have 3 to 5 years of experience. Supervisors typically have many more years of experience.[28]

The contractor appears to have the right mix of personnel available who are sufficiently trained to fulfill their role in the scanning process.

SUMMARY

Based on the information provided by the EPA and the contractor, the Stratus team believes that EPA's digitization practices conforms with recommended library digitization processes. However, this assessment is based purely on anticipated practices conveyed by EPA and its contractor, not on observation of current practices. The contractor appears to conform to recommended processes and is using current and proven technology to scan documents. The contractor described a satisfactory operating environment, a well-trained workforce, and a management staff experienced in processing large volumes of paper. In addition, the EPA contractor's quality control process for checking indexing and resolution appears to be sufficient.

We identified several areas in the proposed digitization process that could be improved, especially in regard to metadata for indexing on keywords and the establishment and collection of metrics on cost, speed, and volume. In addition, we believe that EPA should investigate storing files in PDF/A format, establishing a statistical approach for random sampling, and using HSM for image storage. EPA should also consider establishing a digital archive for long-term preservation of the digital images.

We also recommend that when the digitization of library materials resumes, EPA periodically review the process in operation.

[28] Information provided by the EPA contractor during the interview with the EPA contractor and EPA staff members on August 31, 2007.