



USACE CADD/GIS Technology Center Project 03.036

Quality Assurance Procedures For Historic and Digital Map Collection Methods

January 5, 2004

Prepared For:

Mr. Ralph Scheid
CEMVN-ED-SE
US Army Corps of Engineers, New Orleans District, Attn: ED-SE
P.O. Box 60267, New Orleans, LA 70160-0267
(504) 862-2995 Fax: (240) 525-2497
Ralph.A.Scheid@mvn02.usace.army.mil

And

Mr. Blaise Grden
CEERD-ID
US Army Engineer Research and Development Center
CADD/GIS Technology Center
3909 Halls Ferry Road
Vicksburg, MS 39180-6199
(601) 634-3581
Blaise.G.Grden@erd.c.usace.army.mil

Prepared By:

Contract No:

Task Order Number:

InStep Software, LLC
55 E. Monroe
Chicago, IL 60603

Abstract:

This technical report contains an overview of digital imaging processes, including hardware, software, and storage medium considerations. It outlines best practices and quality control recommendations for long-term storage of digital images, specifically those created from paper maps, engineering drawings, aperture cards, aerial photos, and microfilm, taken from a broad survey of public and professional industry sources.

Table of Contents

Final Technical Report: Quality Assurance Procedures for Historic and Digital Map Collection Methods

1 EXECUTIVE SUMMARY	6
1.1 PROJECT PLANNING AND PREPARATION	6
1.2 CREATING DIGITAL IMAGES	6
1.3 INDEXING AND METADATA	7
1.4 MAINTENANCE AND PRESERVATION	7
2 INTRODUCTION	8
2.1 PROJECT BACKGROUND AND OBJECTIVES	8
2.2 ADCS REPORT SUMMARY	8
2.3 QUALITY CONTROL AND ASSURANCE	9
2.4 DOCUMENT OVERVIEW	9
3 PRINCIPLES OF DATA INTEGRITY	11
3.1 IMAGE INTEGRITY	11
3.2 INFORMATION INTEGRITY	11
4 PROJECT PLANNING AND PREPARATION	12
4.1 SET GOALS BY ASSESSING USERS AND NEEDS	12
4.2 PHYSICAL COLLECTION ANALYSIS	13
4.3 STAFFING REQUIREMENTS	14
4.4 STAFF TRAINING AND EDUCATION	14
4.5 EQUIPMENT ACQUISITION	15
4.6 ORIGINAL DOCUMENT PRESERVATION REQUIREMENTS	15
4.7 ORIGINAL DOCUMENT LOCATION INFORMATION	16
4.8 DIGITAL IMAGE RETENTION	16
5 WORKFLOW	17
5.1 ESTABLISHING QUALITY CONTROL CRITERIA	17
5.2 DETERMINING MASTER IMAGE GOALS	18
5.3 DOCUMENT PREPARATION	19
5.4 SCANNER CALIBRATION	19
5.5 SCANNING THE IMAGE	20
5.6 CREATE DERIVATIVE IMAGES	21
5.7 STORING THE IMAGE	21
5.8 ADDING INDEXING AND METADATA	21
5.9 CREATING A PAPER TRAIL	21
5.10 FINAL VISUAL INSPECTION	22
6 DOCUMENT SURVEY	23
6.1 PURPOSE OF A DOCUMENT SURVEY	23
6.2 IMAGE METRICS	23
6.3 IMAGE METRICS METHODOLOGY	24
6.4 BENCHMARKING AND VERIFYING REQUIREMENTS	25
6.5 PRODUCTION AND WORKLOAD RATES	25
6.6 DOCUMENT FILE STATISTICAL ANALYSIS	25

7 IMAGING PROCESSES	27
7.1 INTRODUCTION	27
7.2 PRE-SCANNING IMAGE ENHANCEMENT AND/OR RESTORATION	27
7.3 DOCUMENT PREPARATION	27
7.4 SCANNING	28
7.5 IMAGE ENHANCEMENT	28
7.6 IMAGE COMPRESSION	29
7.7 IMAGE ANALYSIS	30
7.8 INDEXING	30
7.9 DISTRIBUTION AND TRANSMISSION	31
7.10 QUALITY CONTROL	31
7.11 USER VIEW	32
8 SCANNER EQUIPMENT CONSIDERATIONS	33
8.1 SCANNER SELECTION	33
8.2 FLATBED SCANNERS	33
8.3 SLIDE SCANNERS	34
8.4 DIGITAL CAMERAS	34
8.5 DRUM SCANNERS	34
8.6 HAND-HELD SCANNERS	35
8.7 SCANNER SOFTWARE	35
8.8 SCAN RATES	36
8.9 SCANNER RESOLUTION	36
8.10 NOISE	38
8.11 LIGHT SOURCES	38
8.12 SCANNER SUGGESTIONS BASED ON MATERIAL TYPE	38
9 SCANNER CALIBRATION AND QUALITY CONTROL	40
9.1 IMAGE QUALITY	40
9.2 QUALITY CONTROL ADVANTAGES	40
9.3 QUALITY CONTROL CONSIDERATIONS FOR SCANNER CALIBRATION	40
9.4 SCANNER AND MONITOR CALIBRATION	41
9.5 QUALITY CONTROL REFERENCE TARGETS AND COLOR BARS	42
9.6 WHEN TO RUN QUALITY REFERENCE TESTS	43
9.7 TEST LOGS	44
9.8 CREATING TEST IMAGES	44
9.9 TEST PROCEDURES	44
9.10 VISUAL INSPECTION	45
10 SAMPLE STANDARD TESTS	47
10.1 STANDARD TEST TARGET: IEEE STD 167A SERIES	47
10.2 STANDARD TEST TARGET: AHM SCANNER TARGET	52
10.3 STANDARD TEST TARGET: RIT PROCESS INK GAMUT CHART	57
10.4 STANDARD TEST TARGET: IT8.7	59
10.5 OBTAINING TARGETS	ERROR! BOOKMARK NOT DEFINED.
11 IMAGE TYPES	60
11.1 IMAGE TYPES	60
11.2 DIGITAL MASTER IMAGES	60
11.3 ACCESS IMAGE	63
11.4 THUMBNAIL IMAGE	63
11.5 ENHANCED IMAGES	63
11.6 GRAYSCALE AND HALFTONES	65
12 METADATA	67

12.1 INTRODUCTION TO METADATA	67
12.2 ADVANTAGES OF USING METADATA WITH ARCHIVAL IMAGE DATA	68
12.3 TYPICAL METADATA CATEGORIES	68
12.4 TYPICAL METADATA CHARACTERISTICS	69
12.5 METADATA STRUCTURE	71
12.6 METADATA AND DIGITAL ARCHIVES	72
12.7 METADATA SPECIFICATION OPTIONS	72
12.8 METADATA WORKFLOW	73
12.9 UNIQUE IDENTIFICATION	75
12.10 AUTHENTICATION	75
12.11 SELF-DESCRIBING	76
12.12 INTELLECTUAL PROPERTY	76
12.13 USAGE STATISTICS	76
13 COLLECTING METADATA	78
13.1 ADVANTAGES OF USING METADATA	78
13.2 IMAGE METADATA	79
14 IMAGE FORMATS	81
14.1 A NOTE ON EDITING LOSSY IMAGE FORMATS	81
14.2 TIFF ITU-T.6	81
14.3 JPEG	81
14.4 JFIF (JPEG FILE INTERCHANGE FORMAT)	82
14.5 JPEG 2000	82
14.6 GIF	84
14.7 PNG	84
14.8 PDF	84
14.9 CCITT GROUP 4	84
15 IMAGE FORMAT SELECTION	85
15.1 INTRODUCTION	85
15.2 IMAGE QUALITY	85
15.3 SPECIFIC MINIMUM RESOLUTION AND FILE FORMATS	86
15.4 REFLECTIVE FORMATS	86
15.5 TRANSMISSIVE FORMATS	87
15.6 OVERSIZED ORIGINALS – MAPS AND ENGINEERING DRAWINGS	87
15.7 BIT DEPTH AND COLOR DEPTH	87
15.8 RESOLUTION	89
15.9 IMAGE FORMATS FOR DIGITAL MASTERS	89
15.10 IMAGE FORMATS FOR DERIVATIVE IMAGES	90
15.11 MINIMUM QUALITY LEVEL – PHOTOGRAPHS	91
15.12 MINIMUM QUALITY LEVEL – MAPS AND ENGINEERING DRAWINGS	92
15.13 SUMMARY OF GENERAL RECOMMENDATIONS	93
16 STORING IMAGES	95
16.1 STORING IMAGES	95
16.2 MAGNETIC DISKS	95
16.3 CD-ROM	95
16.4 TAPE	96
16.5 ACCESSIBILITY LEVELS	96
16.6 ACCESS TYPES	97
16.7 RECORDING MECHANISMS	97
16.8 TYPES OF STABILITY FACTORS	97
16.9 PHYSICAL STABILITY FACTORS	98
16.10 DATA STABILITY FACTORS	99
16.11 TECHNOLOGY STABILITY FACTORS	99

16.12 DATA RECORDING AND VERIFICATION	99
17 MAINTENANCE AND PRESERVATION	101
17.1 DIGITAL PRESERVATION STRATEGY: TECHNOLOGY EMULATION	102
17.2 DIGITAL PRESERVATION STRATEGY: TECHNOLOGY MIGRATION	102
17.3 DIGITAL ROSETTA STONE	103
17.4 COSTS	103
17.5 BACKUPS	104
18 GUIDELINES FOR CREATING A REQUEST FOR PROPOSAL	105
18.1 OVERVIEW	105
18.2 ANALYZE FUNCTIONAL REQUIREMENTS	106
18.3 TECHNICAL REQUIREMENTS	106
18.4 PROJECT MANAGEMENT REQUIREMENTS	107
18.5 VENDOR RESPONSE REQUIREMENTS	107
18.6 QUALITY CONTROL REQUIREMENTS	108
APPENDIX A. REFERENCES	110
APPENDIX B. DEFINITIONS	113
APPENDIX C. EXAMPLE RFP	115
APPENDIX D. SPECIFICATION FOR QUALITY CONTROL AND METADATA BUILDING TOOL FOR MANAGING SCANNING PROJECTS	161

1 Executive Summary

Any successful digitization project requires proper planning and the establishment of quality assurance procedures. The development of the project's goals is important since the purpose behind digital imaging may be either preservation, enhancing access, or a combination of these purposes. For preservation purposes it is important to create a high-quality scan because the source document may only be available for conversion once and the digital image must serve as a surrogate. For enhanced access the emphasis may be on obtaining derivatives suitable for supporting user's needs for printing and display, and input into various software systems. For derivatives completeness and detail may vie with speed of output requirements. User expectations may become more demanding over time, so even if current requirements dictate a reduced image size a digital master should be rich enough to accommodate future derivatives of greater detail. Since no two digitization projects are identical each project must identify the minimum level of quality and information density it requires for its digital surrogates.

Quality digital images are obtained by following an established quality assurance plan throughout the entire life cycle of a project. Stages to be considered in quality assurance development for a digitization project include: Project Planning and Preparation, Creating Digital Images, Indexing and Metadata, and Maintenance and Preservation.

1.1 Project Planning and Preparation

Quality control begins with project preparation. Due consideration should be given to project planning at the beginning of any digitization initiative and reviewed any time there is a change in the overall goals or needs of the project, or when new equipment is to be procured.

Quality is a subjective term; for a digitization project a quality image is defined by how useful the image is to the end user. Therefore, quality assurance starts with understanding what is needed by the end users of the digital images and what type of documents exist in the physical collection.

A document survey will help quantify what types of originals will need to be scanned and can be useful in specifying scanner selection criteria (if new equipment is to be procured) and planning how to best scan the existing collection.

Finally, project objectives should be clearly and explicitly defined, then communicated to all project staff. Project team members will be the implementers of quality control; know what the project goals are will help them determine what represent a quality image and what does not.

After adequate training is given to project staff members, scanning documents and creating digital images can commence.

1.2 Creating Digital Images

The scanner has the greatest impact on image quality, so verifying the accuracy of scanning equipment is essential for successful quality control. Scanner accuracy can be verified using a series of standard test targets, sample documents (taken from the document survey), and proper visual inspection.

Once the scanner calibration has been verified and scanning commences, visual inspection and comparison will further assure that quality images are created.

The image created from the initial scan should be of the highest applicable quality - based on the project needs - and stored in a lossless image format. This image file is called the digital master image.

Subsequent derivative images will then be created from this single master image. Most commonly, two derived images are will be made: a thumbnail image and an access image.

Thumbnail images are small, low resolution images used to “preview” the image, typically in a web page. An access image is the image that most users will work with. Derived images can be stored using a lossy format (meaning a file format that loses some of the image information to achieve better compression and smaller file sizes),

Additional “enhanced” images can also be created to improve contrast, remove background noise, deskew the images, etc. The enhanced images are another type of derived image, and should not replace the master image.

The images are then saved to some permanent media (e.g. CD-ROM, network drive, etc.).

1.3 Indexing and Metadata

Metadata, which is data about data, is used to describe the contents of a digital image. There are a variety of different metadata styles that can be used, but all generally provide a structure for storing information about the contents of a digital image, how it was created, what version it is, and other useful data.

Standard vocabularies, which will ensure that metadata terms are used consistently in descriptions, will further increase the ease with which metadata records can be searched.

Indexing is the process of saving the location of digital images within a file system in a searchable database.

When indexing is linked with metadata, a powerful tool for finding images and related documents becomes available to users.

1.4 Maintenance and Preservation

Long term digital archive projects all require a certain level of maintenance to preserve the usability of images. This ongoing effort and its associated costs should be accounted for as a fundamental part of any digitization process.

Back up copies of all important files (e.g. the digital master, metadata, and indexing database) should be made and stored both on-site (for short term recovery) and off-site (for long-term) recovery.

Because all media suffer degradation over time, new copies may need to be made from the original media on a periodic basis. Software solutions exist to ensure that copies are created accurately.

As technology changes, image formats, software, and hardware become obsolete. Images stored using an obsolete format or on obsolete hardware will be inaccessible, and the recover costs may be too high to make them available again.

This risk can be mitigated by using a combination of technology emulation and migration. Emulation allows older hardware or software to be mimicked by current systems. Migration is the process of transferring data from an older system to a more modern one.

2 Introduction

2.1 Project Background and Objectives

This document compiles and establishes standardized scanning specifications and quality assurance tools for scanning existing documents to create digital images in an acceptable quality and format. It is part of a larger CADD/GIS Technology Center R&D Work Unit 03.036, “Historical and Digital Map Collection Methods and Online Retrieval Tools.”

Specifically, this document covers quality assurance and control “best practices” for establishing and maintaining a digital image archive. Quality control includes the procedures and practices that are put in place to ensure the consistency, integrity and reliability of the digitization process. Quality assurance refers to the procedures by which the quality of the final product is checked. The information contained herein was drawn from a variety of sources such as the National Archives and Records Administration, various digital library initiatives, IEEE¹, the Library of Congress, and other professional standards boards. As such, it represents a survey summary of current industry and public institution recommendations.

Special attention has been paid to the longevity of the data, its projected usage, and portability of file formats for maintaining quality of both raster information and associated pure data. At the time of this study, the primary focus of the USACE image digitization effort was centered on the following media types:

- paper maps (both historic and working)
- engineering drawings
- aperture cards
- aerial photos
- microfilm
- related documents

The current document provides the methodology and technical information necessary to perform the imaging, emphasizing the quality control and quality assurance tracking procedures required during the digitization process.

2.2 ADCS Report Summary

This document also provides follow-on study and supplement to a previous report entitled: "Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative"². This previous report presented benefits of digitization, including increased access, lower retrieval time, sharing of document collection, a streamlined workflow, and reduced storage costs. It surveyed three representative sites, concluding that large volume was the overriding impediment to any digitization efforts in the Corps. It concluded that there needed to be better storage mechanisms than CD-ROM libraries and presented a software application to aid in the loading of previously scanned documents to one such system (Bentley’s Digital Interplot) with the Metadata Collection and Bulk Loading System (MCABLS).

The study showed that the existing USACE hardcopy library is extensive, and inadequately indexed, which can make finding documents time consuming or impossible. Improving document access is of critical importance to the US Army Corp of Engineers; the study found that implementing document digitization would be the easiest way to increase accessibility. While many USACE departments have begun to digitize their records there is

¹ Institute of Electrical and Electronics Engineers. <http://www.ieee.org/portal/index.jsp>

² Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative – Final Report”, Contract Number: N66032-94-D-0012, Intergraph Solutions Group, http://tsc.wes.army.mil/downloads/ADCS_Final_Report_Main.pdf.

currently no standard for scanning or associating metadata. Many departments are not using metadata at all, which makes it difficult to search the digital records. The USACE needs to establish a set of standards for collecting, scanning, indexing, and associating metadata with digital records. One recommended metadata structure cited in the study is that supported by ARIMS³.

Additionally, the study noted that many documents are deteriorating but that recreating information lost in deteriorated documents is costly and sometimes impossible. The USACE is, therefore, slowly irrevocably losing some of the information in its archives to deterioration.

The ADCS (Automated Document Conversion Strategy) Initiative study showed that if a move to digital records was implemented, it would provide faster, more reliable document access at a lower cost while at the same time reducing labor and storage costs. According to the study, the benefits of implementing a standardized digitization effort (and conversion) should far outweigh the initial overhead costs.

2.3 Quality Control and Assurance

For any image digitization project, the focus of quality control and assurance centers on preserving data integrity. Data integrity refers to the long-term preservation of both the visual representation of the digital image and the data associated with it. A more in-depth discussion of data integrity is presented in Section 2.

Following project planning activities, digitization for any image has three main phases: scanning of the original physical object (the “hardcopy” image), associating additional information with the electronic image, and storing the results. Each of these phases has been given individual consideration. The overall work flow has also been considered in establishing the QA/QC procedures outlined on the following pages.

Maintenance must also be considered a fundamental part any digital image initiative, if long term usage is expected. As such, quality assurance and control standards for long-term maintenance of digital imaging data have also been included in this survey.

Simply put, strict quality control ensures that data images can be located and that the data stored is of an acceptable quality level.

2.4 Document Overview

This document follows the steps required in any digitization project, from project planning through the imaging process, to indexing the documents’ metadata, to storing the images. It first outlines the main principles of data integrity associated with a digitization initiative, which will provide the foundation for establishing quality control criteria, and then explores the general tasks and workflow associated with digitization. Essentially, every point in the workflow represents a potential point of information loss, and therefore has been explored for quality control and assurance opportunities.

³ Army Records Information Management System. A record management system available to the USACE that provides an indexing and metadata database, as well as access control to digital records.

This document presents a categorization of the results of the aforementioned survey of quality control measures in the following main topic areas:

Topic	Section
Project Preparation	4,5,6
Scanning and Image Processing	7,8,9,10,11
Associating Data with Image	12,13
Storing and Maintaining the Image	14,15,16,17
Guidelines for writing an RFP	18

The Appendices present references, definitions, excerpts from a sample RFP, and the specification for the software tool to be used during the scanning process. This tool, the “Quality Control and Metadata Building Tool” (QCMBT) ties together the quality control and assurance processes detailed here to ensure that scanning, metadata collection, and storage are an integral process.

3 Principles of Data Integrity

Digital objects can be any type of record or information represented in the digital world. In the simplest case, a digital object might be a single image file. In a more complex case, the digital object might include an image file, references into a database and descriptive information about the contents of image, including how it was created.

The information in a good digital object should be persistent. Yet digital information is notoriously volatile. Because of this volatility, digital objects are not likely to persist without the specific intention of some individual or institution that ensures the information in the digital object will persist. It will require an applied effort to ensure that the digital object remains accessible over time despite changing technologies.

In the world of traditional preservation, the goal is to preserve the physical object from physical degradation. In the digital world, the goal becomes to preserve the intellectual integrity of the digital object from degrading. The intellectual integrity of the digital image representation of a physical record is comprised of two main areas: the image integrity and information integrity associated with the record. An understanding of these terms will be useful in later sections when issues related to subjective estimates of image quality will need to be assessed by digitization project managers.

3.1 Image Integrity

Image integrity can be thought of as how accurately the digital image reflects the physical object. There can be some loss of information when converting from an analogue source (physical object) to a digital representation (scanned image) depending on the resolution of the digital image. However, there are also some practical limits on what level of resolution is appropriate for any specific task.

Web-based thumbnails, for example, don't need to have the resolution of the original image. Even a digital master, which might be used to make enhanced enlargements, may encounter practical limits, such as the granularity of photographic images.

Resolution isn't the only factor that affects an image's integrity. Color accuracy and depth, and optical distortion are also important contributing factors that will effect how accurately the digital image reflects its physical counterpart.

3.2 Information Integrity

Information integrity refers to additional information associated with the digital image. Sometimes the information can be recorded with the digital image, such as a legend imbedded in a map; sometimes it may come for an external source, such as a catalogue number. Information can also change over the lifetime of the digital image.

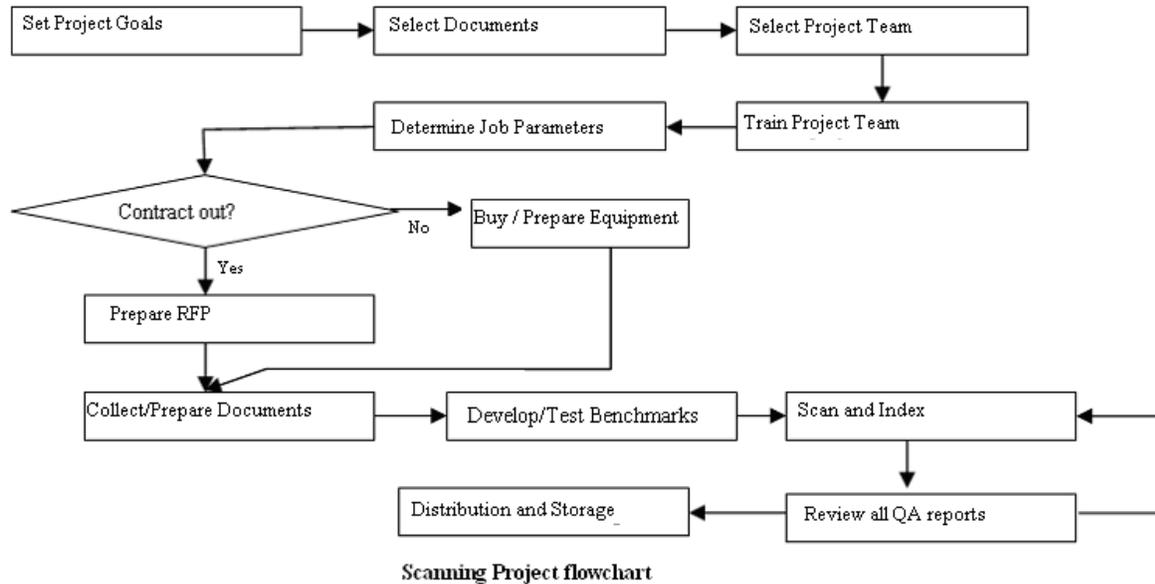
The information might also be comprised from many different sources (e.g. indexing database, scanning information, etc.). In some cases this information can be stored as part of the image file itself; several image formats provide space in the image header or additional information about the image file that are not directly related to the rendering of the image. In other cases, the data is stored externally.

Regardless of where the information is stored, preserving the integrity of that information means not only preserving the actual data itself, but also the reference back to the associated digital image. Preserving the referential integrity of the information means that, for example, if the digital image is moved or renamed the information associated with it will be modified to point at the new file.

Good information integrity will help users determine what an image is; and, if needed, authenticate that it is what it claims to be.

4 Project Planning and Preparation

Quality control starts with proper project planning and preparation. The areas outlined in this section should be considered at the beginning of any digitization initiative. They should also be included (or at least reviewed) any time there is a change in the overall goals or needs of the project, or when new equipment is to be procured. The following flowchart shows an overview of the steps in a typical digitization project. The first several steps are discussed in this section, the latter steps are discussed throughout the rest of this document.



4.1 Set Goals by Assessing Users and Needs

Goals for a digitization project should be documented in an **explicit collection development policy** that has been agreed upon and documented before digitization begins. Before starting any imaging project, know the project's mission, users, priorities (speed, image quality, and quantity), and functional goals (reference, web use, publication, other). Project managers should be able to identify the target audience(s) for the collection (both in the short term and in the future) and how the selected materials relate to their audience. If the materials exist in non-digital form, their usage statistics should be examined, if they are available. Special attention should be given to any factors that will influence the use or value of documents when they become digitized.

If the way that documents will (or could) be used isn't carefully considered during project planning, the implementation of the project risks reducing the use or value of the resultant digital documents. The format used to store digital images and their associated information (indexing and metadata) should be chosen based on both the intended current and likely future use. At the very least they should support the development of access copies that support those uses.

One of the primary motivations for any digitization effort is usually to increase the availability of documents in the collection. This offers increased accessibility for the end users and reduces degradation of the physical originals because they will no longer have to be handled physically to be accessed. Consequently, records in a good digital project will be broadly accessible. In almost every case, there is a direct correlation between the production quality of a digitized object and the readiness and flexibility with which that object may be migrated

across platforms. As a result, the digitization of objects at the highest affordable quality can pay off in the long run as the objects are rendered more useful and more flexibly accessible over the longer term. Thus, due consideration should be given to the required longevity of digital images

Imaging requirements, such as resolution, compression, headers, will vary depending on how images will be used [5]. What constitutes a "good image" for the purposes of your project--a faithful reproduction (no data loss) or a pleasing image (a view graphic or altered image)--should be defined prior to scanning. This is the difference between a project whose primary goal is preservation or accessibility/usability. The initial examination of how images will be used should also define the specific indexing requirements and metadata fields that will be used by the project.

Consideration should also be given to user equipment. Consider how users will connect to the image sources (through a dial-up account, LAN, directly from CD) and what type of viewing hardware they will use. The size of the monitor as well as the resolution and color palate of the typical user should be considered. These values can be used to determine the parameters that should be used for access images. See Section 11: Image Types for more information on access images.

Consider how the particular digital collection will fit in with the organization's overall document management policy, as digital collections should not stand in isolation from the original materials or from the policies of the organization as a whole. Record keeping requirements specific to the collection should be considered, and developed where necessary, based on the specific business needs for the system, agency IT architecture, and overall project mission. An explicit collection policy should address the treatment of archival files, working file and backup files (including storage and materials). Asset management should also be considered, along with file safety, protection against hazardous materials, water damage, and theft. This includes determining how digital images are handled, converted, stored, disseminated, and secured. Environmental and local building codes for air conditioning, electrical power, ventilation, and location of equipment should be examined and adhered to for any project.

Software standards and conventions for custom software, and functional and performance characteristics of commercial off-the-shelf software (COTS) should be well documented. This also follows for hardware standards according to FCC, Underwriter Lab, etc. plus standards for reliability and maintainability. Quality standards should be assigned to assure uniformity and compliance with standards codes, policies and regulations (ISO⁴ 9000, ANSI, IEEE, FIPS Pubs, DOD Instructions, OMB, local safety and fire codes, others) governing how digital images are handled within a project. The digital project may wish to consider implementing records management policies used by other archival projects, particularly if the project expects to interoperate with those systems. For example, the NARA⁵ approach to records management is based on the ISO Records Management Standard 15489.

The protocol put forth in this document is drawn from various AIIM International recommendations, NARA standards, National Information Standards Organization (NISO) releases, MIL-STD1840C and MilSpec_28002C, as well as compatibility with the Corps' own standards for electronic document management systems.

4.2 Physical Collection Analysis

Different document types will work better with some scanner types than with others. The volume of the collection, when combined with a summary of the needs of the users will determine the minimal initial storage requirements. Future needs can be extrapolated from growth information about the physical collection.

For more information on performing a document survey on an existing collection, see Section 6.

⁴ International Standards Organization <http://www.itu.int/home/>

⁵ U.S. National Archives and Records Administration <http://www.archives.gov/>

The best situation is one where the source materials and project goals dictate the image quality settings and the hardware and software one employs. Digitization projects should base the decision about a document management system's hardware and software on the organization's original paper documents.

4.3 Staffing Requirements

Before starting a project assess staff expertise and availability (to do scanning, manage infrastructure, migrate data, and build metadata), and address content issues, such as physical condition, format, nature and attributes to be captured. Any digitization effort will require a combination of staff with different areas of expertise. The nature of a digitization project is such that it requires a team approach. The following areas and skills may be important to any digitization project:

- Project management skills
- Technical skills/staff
- Database development and administration skills
- Cataloging staff/skills
- Computer programming skills
- Web design skills
- Subject matter specialists
- Knowledgeable end users
- Preservation background
- Photography background
- Artistic/graphic design skills

These groups should be consulted for input on project specifics in their areas as well as input on staffing and labor requirements.

Some digital imaging projects may not have dedicated staff working on the project; instead they will utilize existing staff from other areas in the organization or augment the existing staff with outsourcing. It may benefit the project to look at "transferable skills" that staff members may already possess that would be useful in any digitization project. Sufficient time for training, and opportunities to receive further education and training, should also be provided.

Increased imaging and indexing of records and quality control procedures may require additional staff training. Projects may also incur migration costs if the information has to be retained for periods longer than five to ten years. Staffing should be carefully considered. If workload requirements exceed staff workload capacity, quality control tasks may be given lower priorities in order to meet demands. The same risk applies to maintenance tasks. Both of these areas represent potential areas for information loss.

Within the Corps of Engineers, use of a Project Delivery Team (PDT) in implementing the Project Management approach to a scanning effort will add quality and established processes to the effort.

4.4 Staff Training and Education

Educating project staff and administrators about the issues involved in successfully planning, implementing, and maintaining a digital collection is extremely important. Staff should be aware of potential areas where information can be lost, and their responsibilities for maintaining data integrity.

Additional time must be allocated for training with specific scanning hardware and software, as well as indexing and metadata data-entry software, systems back-up and storage. If new equipment is acquired during the course of the project, additional time should be allocated for training on the new systems.

In addition to providing training, staff members should be educated about the goals of the project. When staff members are aware of how the digital records will be used and what will make them valuable and useful to the end clients, they will be able to make better judgments about what constitutes a good, “quality” image.

Because of the subjective nature of assessing the quality of an image, it is essential that these goals be communicated clearly for the onset of any digitization project.

4.5 Equipment Acquisition

The quality of digital images produced by scanning is directly related to the specific hardware and software used. The specific documents can be determined from a document survey of the existing collection.

There can be a big difference between scanners. That difference is not just limited to high-level concepts like “resolution”, and “scan rates”. Scanners also vary in how they treat specific documents and images. Some scanners may not have a constant (i.e. linear) performance across an entire image, or may have complex interdependencies when processing an image. Moiré patterns are a good example of this type of inconsistency; even if a single image is used, the moiré pattern created by different scanners is likely to be different in the resultant digital images.

Knowing roughly what types of images will be processed, and ideally having a sample collection available during the evaluation, will help ensure that a good combination of hardware and software can be procured.

There are approximately 15 different scanner classes [20] operating at different and mixed resolution ranges, processing speeds, tonal quality, formats, enhancement capabilities, dealing with bi-tonal, gray scale, and color imagery on paper or film exclusively or in combination.

These scanners will process different document types differently, and create different types of file output formats. There are about 50 different formats, many of which are proprietary and not necessarily universally transferable. Scanner output has to be synchronized with about a dozen different types of appropriate compression algorithms, such as CCITT Group IV, JPEG, JPEG2000, wavelet, fractal, among others; plus techniques as lossy (may be some loss of data on the original image with images not precisely constructed back to the original image form) or lossless (exact reconstruction/replication of image pixels). The right selection of algorithm depends on the resolution and format produced by the scanner.

Depending on the size and composition of the documents in the collection, different scanners may be employed to scan different document types. For example one type of scanner might be used to scan text, while another is used for oversize items, photographic prints, slides, or other formats. Some scanners will work better with some document formats than others.

Scanner and compression capabilities are to be synchronized with different image enhancement capabilities (dynamic thresholding, rotation, edge detection, segmentation, scale to gray, among 120 other algorithms) for different types of document resolutions, formats, and compression produced.

In addition to all of the file formats and scanning software, there are perhaps more than a dozen [20] different types of individual optical character recognition engines. The accuracy of the character recognition systems is closely tied to the scanning, compression, enhancement and storage format of the digital image. Ensuring that these factors are properly balanced for good character recognition can be another challenge and should be considered if that that will be part of normal scanning workflow.

4.6 Original Document Preservation Requirements

In some cases digitization is performed in order to reduce the volume of physical documents. Often times, after a certain period of time, the original hardcopies of digitized images are destroyed. Some consideration should be given to potential future requirements of original documents.

Consider, also, that some documents [5] have information which becomes invalid when transferred to a digital medium. As an example, consider that forensic analysis of certain artifacts, such as signatures, is not possible with imaged records. If your collection contains documents that must also be preserved in their original form, this should be identified and communicated to staff.

4.7 Original Document Location Information

If future digital images are to be created from the original, it will be vital to locate the original physical document. The location of the physical document should be recorded in the indexing information associated with the digital image. The indexing should be able to track the current physical location, past physical locations, and the date/time, method and reason for transfer between physical repositories.

There is some possibility that the USACE and CAD/GIS Center partner's data will be transferred to a Records Holding area or to NARA. The indexing database should account for tracking hardcopy or scanned information that has been transferred another repository, so the indexing should also accommodate the acquisition or import of documents that are external to the project.

This information can be useful in locating the actual physical object, and can provide supporting information for restoration considerations or other future project planning.

4.8 Digital Image Retention

The retention period for any given type of document should be clearly stated in the project plan. Are the records of an important historical nature, which should be preserved indefinitely? Are they purely administrative records that can be destroyed after some appropriate period?

Clearly stating the retention goals will help projects plan for future tasks like storage and technology migration, which helps ensure that the project will be able to retain the records for the required period.

Consider, also, that technology is quickly evolving. What is regarded as a "high resolution" 1200 dpi⁶ file today may be considered unsatisfactory in a few years. Imaging software will continue to evolve and improve. The same can be expected of storage capabilities.

The way in which data is presented to the user might also change. For example, few users accessed data through a web browser (or even knew about the internet) in the early 1990s, but it is now a common method for communicating information.

If the expected useful lifetime of a digital object is well known, project planners can determine if migration to new technology standards is warranted and determine when the migration should take place. Retention planning also lets project planners balance storage cost and capacity with index, conversion, quality control, and migration costs so that resources are not misallocated.

Finally, storing the retention data in the metadata will also help manage the digital objects and create a record of the intended retention period that can "travel with the digital image".

⁶ Dots per inch. The number of "dots", per linear inch used to print a picture. The larger the number of dots per inch, the smaller the dots can be, and therefore the sharper the printed picture. The equivalent for digital images is PPI (pixels per inch).

5 Workflow

Quality control is an ongoing process. As such, it should be an integral part of a well-organized workflow. The process and procedures for handling images, collecting metadata, and cataloging results will be crucial to the success of any digital imaging initiative.

5.1 Establishing Quality Control Criteria

A good workflow is essential for producing reproducible, cost-effective results from the day-to-day operation of a digitization project. Good quality control, with specific standards, will help ensure that the product of the workflow is a quality digital collection. The quality control principles, in turn, should be clearly stated based on the number and type of documents in the collection and the project's goals. Image metrics can be determined by using standard image tests, such as those outlined in sections 9: Scanner Calibration and Quality Control and 10: Sample Standard Tests. Elements of these procedures should be noted in the metadata associated with each digital image.

The workflow process should be well documented, including project principles, imaging procedures, and explicit rules for naming and the vocabulary that will be used for creating metadata. It should include an explicit set of quality assurance and quality control tasks.

Quality assurance encompasses all of the planned actions designed to provide adequate confidence that the images produced by the digitization project will at least meet the minimum required quality level. Quality assurance procedures and goals should be defined that will encompass all of the planned tasks that will be followed in order to be confident that the scanned images will be of a sufficient quality for their intended purpose.

Quality control represents the individual standards and metrics for individual actions within the work process whereby quality levels can be maintained. Individual quality control procedures must be instituted both while preparing documents for imaging and while verifying and validating imaged information that will set specific quality levels. For images the quality levels will be related to the quality of the image as compared to project specific standards; for documentation the quality control will be related to the type of information contained within the documents themselves.

The goals of the project and the needs of the end users determine how the quality of an image should be judged. The determination of an image's quality therefore ultimately depends on an understanding of who are the users (and potential users), and what kind of uses will they make of this material. If not enough quality (in terms of resolution and/or bit-depth) is captured during the initial scanning some potential use will be inhibited. The same is true of the information associated with the digital image. Enough contextual and supporting information must also be added to the record to make sure it meets the usability need of the end users.

Understanding these needs requires that quality inspection personnel have a good understanding of the project's goals and objectives.

Image quality depends on the project's planning choices and implementation. Project designers need to consider what standard practices they will follow for input resolution and bit depth, layout and cropping, image capture metric (including color management), and the particular features of the capture device and its software. Benchmarking quality for any given type of source material can help one select appropriate image quality parameters that capture just the amount of information needed from the source material for eventual use and display. See Section 10: Sample Standard Tests for some industry standard samples of quality control references.

Digital quality control, metadata capture and management, and image capture and management are complex and time-consuming processes requiring expertise and constant vigilance.

5.2 Determining Master Image Goals

The file created by the scanning process, called the digital master image, will be the closest image to the original document. This image will be used to create all the other images associated with the original document (e.g. thumbnail images for view over the web, access images for day-to-day usage, and enhanced images to meet specific image demands). Therefore, it should be of the highest possible quality, which will help ensure the ongoing value of the digitization effort, and ease the process of creating derivative files.

In order to create the best master possible, the goals of the master need to be carefully considered, so that criteria can be established for determine what constitutes a quality master image.

Image quality for digital capture from originals is a measure of the completeness and the accuracy of the capture of the visual information in the original. There is some subjectivity involved in determining completeness and accuracy. Sometimes the subjectivity relates to what is actually being captured. For example, with a manuscript, it may only be important to capture the writing, or it may also be important to capture the watermark and paper grain important as well. At other times the subjectivity relates to how the informational content of what is captured will be used.

For example, consider an original hardcopy that shows faded or stained handwriting, as might be present on some of the older USACE hand made maps. There are two possible approaches one might take. Should the digital representation be artificially enhanced to improve the legibility of the handwriting, or should the digital image reflect the illegibility of the source material? In some sense, both approaches represent a “loss” of information. Preserving the illegible handwriting just transfers the “lost” information of the original to the digital realm, whereas artificially changing the image represents “losing” the information the handwriting was faded. This can be a common problem with many media types that degrade over time, especially if the samples are relatively old.

The figure below shows the original on the left and the scanned image on the right. The one on the right, scanned in grayscale, may appear clearer, but clearly there is loss of information about the color of the original. If a faithful representation is desired, the scan on the left is more accurate; whereas the scan on the right may provide more usability because of its higher clarity.



Figure 5-1 Original (left) and Grayscale (right) Images

Similarly, the figure below shows the original photograph on the left and an artificially enhanced image on the right. The original photograph exhibits a color cast that is the result of the film that was used, but is a faithful representation of the original document. The image on the right uses image correct to remove the color cast, and is a representation of the *intent* of the original document.



Figure 5-2 Original (left) and Enhanced (right)

If the digital image is made to look “better” than the original, what conflicts does that cause when a user comes to see the original and it looks “worse” than the onscreen version?

Projects can help eliminate these conflicts by using both a master image and enhanced image and linking them with the appropriate indexing and metadata. But in order to be assured that the relationship is created properly, clear quality control procedures should be put in place.

5.3 Document Preparation

Determine how much work needs to be done to make the files ready for imaging. For large collections, document preparation often requires a significant work effort. See Section 6: Document Survey for more information.

Some scanners may require adjustments depending on the size or type of document to be scanned. To minimize readjustments, documents should be grouped into like types (based on scanner adjustment requirements). The appropriate groups can also be determined from the document survey.

5.4 Scanner Calibration

The scanner should be calibrated to ensure that the digital images created are as accurate as possible. The specifics of how an individual scanner should be calibrated will vary from model to model. In some cases specialized technicians may be required to service and recalibrate the scanner.

Even if the personnel doing the scanning will not be able to change the scanner calibration, they should perform calibration tests to make sure that the scanner is properly calibrated. These tests involve scanning one or more sample targets and verifying the accuracy of the digital images created.

The test should be run at the beginning and end of every work shift and at the beginning and end of processing for each batch of like documents.

See Section 9: Scanner Calibration and Quality Control for more information on running calibration tests.

5.5 Scanning the Image

Scanning the image is the heart of converting physical documents to digital images. Although it is not the first point in the quality control chain, it is the place where day-to-day operations can have the greatest effect on the variance of quality. As such, it should be subject to more quality control procedures than any other point in the workflow.

Scanning includes the loading of images, running the actual scanning hardware and software, unloading the image, and making judgments about the quality of the resultant image.

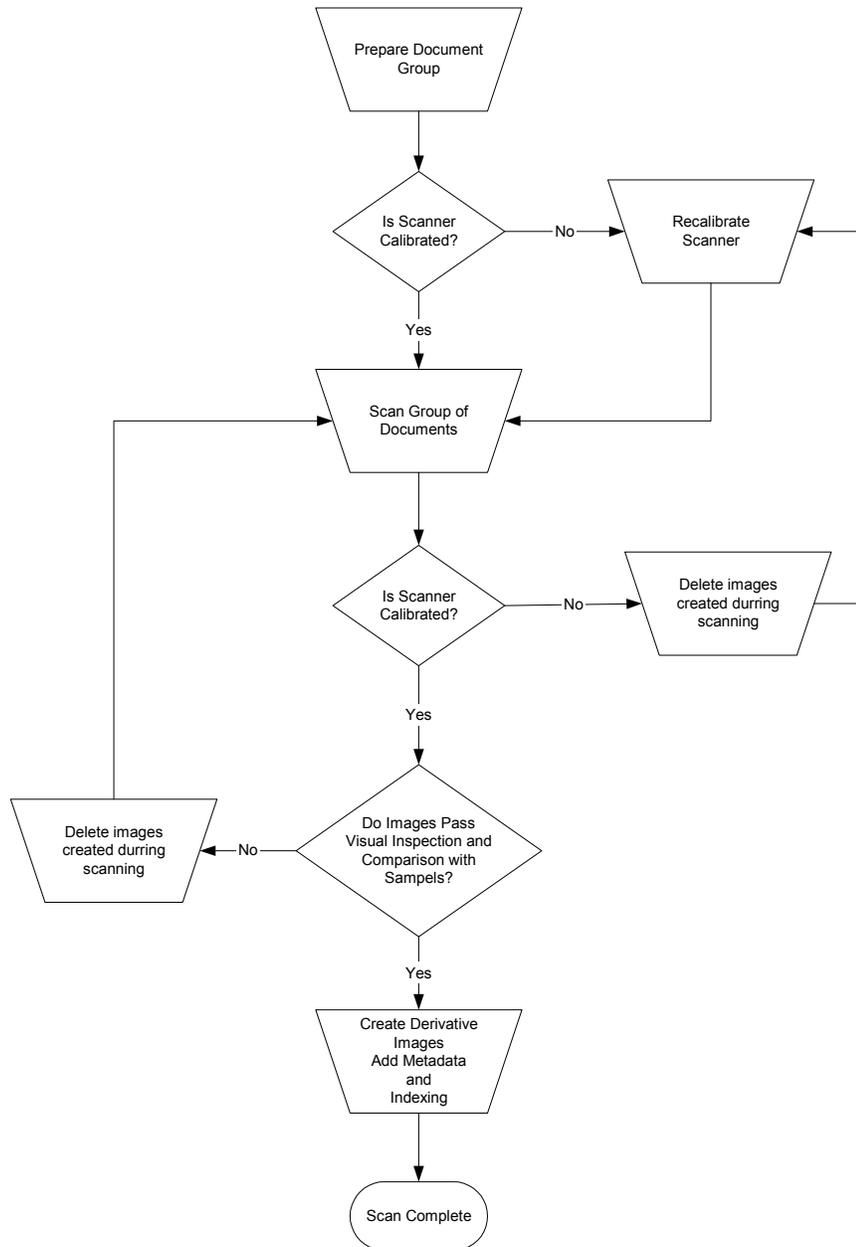


Figure 5-3 General Scanning Workflow

5.6 Create Derivative Images

Derivative images are images created from the digital master. They typically include a thumbnail image, browsing access over the network or through the web, and an access image that is sized and scaled to work well with end user equipment. Digital enhancements or any other images created from the original digital master all fall under the category of derivative images.

Generally, projects will have a standard set of derived images that should always be created for all digital masters. It is a good idea to create these at the same time the digital masters are created, so that they will be available for immediate association through metadata, although they can be created and added at any point after scanning is completed.

5.7 Storing the Image

This is the transfer of the digital image from short-term storage (scanner or computer memory) to mid and long-term storage (hard drive, tape, CD-ROM, etc.).

Verify that the images transferred correctly and can be accessed from the new media. If they are inaccessible, new media may need to be obtained or new writing hardware/software. The data can then be re-written to new media.

File verification should also be done if the image comes from an outside source, as there may be errors on the disk (or whatever transport medium is used) and it may need to be sent back. The results of the file verification should be added to the paper trail, both physical and digital.

5.8 Adding Indexing and Metadata

Considerations should be given to how records are routed, how information is added to records or files, and when records (finals or drafts) need to be captured. This should include unique identification information, location information about where the image is stored, and basic scanning information, at a minimum.

If the data comes from an outside source, such as another project or outsourcing contractor, this information should also be noted in the indexing associated with the location and source of the storage media.

Any non-expert metadata that is available should be added. The project goals should clearly state what types of information can be added after the initial scan. Providing personnel who will do the initial metadata entry with a standard vocabulary and the associated training will help ensure that metadata is created at constant quality level.

Expert metadata can be entered at a later date. The project plan should spell out time frame expectations for the inclusion of expert-level data.

5.9 Creating a Paper Trail

Following established procedures and maintaining the documentation of audit trails and other business practices will ensure that information is kept that may be needed to document record authenticity and reliability. Quality control information and scanner calibration settings should be logged with both the hardware equipment and the image itself.

For more information on scanner calibration, see section 9: Scanner Calibration and Quality Control.

Ideally, a “digital” paper trail will also be created and associated with the metadata of the image. This information should duplicate the information in the physical paper trail. It will track the source of the data, the physical location, backups, whether or not the image was resubmitted, etc.

5.10 Final Visual Inspection

Before being released to the end users, it is generally recommend that a final visual inspection be done on the images. Some digital archivists [38] recommend that the inspection be done by someone other than the person who created the initial digital image.

The final inspection should make sure that the image is appropriate for the needs of the project and look for any unacceptable defects in the digital image.

6 Document Survey

6.1 Purpose of a Document Survey

There are a wide variety of visual elements that may be present in the original hardcopy collection of documents that may require different scanner settings (or different scanners) in order to create accurate digital images.

“The constant case is that each and every single image is a completely independent (virtual) object, carrying its own message and unique physical features, different from every other document in the world. For instance, an original and its duplicate are in reality truly separate and different documents with different legal definitions and different imaging qualities, as well. Image metrics methods preserve these properties and definitions through proper image processing technologies.” [20]

A document survey consists of analyzing some or all of the physical documents in a collection that are under consideration for digitization. The survey is meant to identify which types of document groups are present in the document collection. The individual groups are based on the physical and visual composition of the elements as related to scanning.

The survey is normally used in one of two ways. Firstly, it can be used to create a statistical summary of what types of documents need to be scanned. This information can then be used to determine what types of scanners should be used for the digitization and can help identify if scanners with new capabilities will be required.

Secondly, it can help determine what scanner configurations should be used to best capture the documents.

The set of “test” documents should contain at least one representative image from each image group. This means a “typical” image, usually selected at random, that has roughly the same color and tonal balance as other images in the group. Additionally, the test group should include images that represent the “extremes” of the image group, such as those with high concentrations of one color or hue. The number of documents in the “test” set will be based on the diversity of all available images, the number of image groups in the full document archive, and resources of the group doing the document survey.

This set will be used to quantify metrics, perform benchmarking, verify requirements, establish workflow rates, and test metadata. The documents can then be added to the collection of test targets, which will provide additional resources to project staff when making determinations about digital image quality.

See section 10: Sample Standard Tests for more information about test targets.

6.2 Image Metrics

Image metrics are a way of identifying the characteristics of the physical source documents in a collection. These metrics can then be used to define image processing requirements. These characteristics or requirements can be categorized into four types:

DOCUMENT IMAGE PROPERTY

The nature of each original hardcopy image and its intrinsic physical properties are vital for appraisal. The document’s message or its meaning is not directly important for selecting image processing technologies (except perhaps to assert priorities for workflow processing).

The document image property is based on its physical characteristics. The type of medium on which the document is stored (e.g. paper, film-based or other material such as plastics) is generally the most influential

property, and can form the basis for grouping images if resources can not be devoted to further classifying documents in the collection.

The image property is also related to other visual qualities that apply to the whole document, such as glossy finish and whether the surface is reflective or transmissive.

Some visual elements to watch for include: free text, film products, and emblems and other graphic elements. Clarity, shape, density, wholeness, contrasts, and legibility are other characteristics that should be noted.

When OCR is a consideration, free text correspondence which can be a combination of free printed machine text plus graphics which could require different imaging technologies to replicate; free text could also be represented by unconstrained handprint text.

Film products, such as positive and negative photography, microfiche, roll film, slides, photographs, or large format transparencies may require special processing, due to reflective/transmissive nature of the surfaces, and the potential for optical reflection from both the top and bottom edges of the document. Typically this involves controlling the light source to minimize reflection or placing the document under a transparent non-reflecting surface.

Illustrated emblems or graphics, especially maps, generally have a different composition than text because of colors, backgrounds, and on-linear arrangements.

IMAGE CONTENT ARTIFACTS

Content artifacts are simply the visual components of the image (e.g. text, graphics, pictures, line art, noise, gray scale, color, black and white, etc.) that can be used as image markers.

Content can also be used to describe the physical material on which the image is stored and its condition. Glossy images may require glare compensation. Delicate records may need to be placed in transparent cases before scanning. In general, the condition of the records will affect both the handling during imaging as well as the quality of the imaged record that can be produced and the types of scanners that should be used. This will particularly be a factor for records that are damaged, faded, and/or oversized.

REQUIRED QUALITY

Properties such as resolution needed, tonal quality, sharpness, sizing, focus, etc. constitute the required quality level. These should be based on the explicit quality requirements defined in the project specification.

PLANNED IMAGE USE

Planned image use refers to how the data will be accessed or made available. For example, will the document be printed, or transmitted over the network? Other examples include archiving, on-line viewing, and data capture.

6.3 Image Metrics Methodology

The aim of the metric analysis is one of statistical discovery. The aim of the analysis is to produce a count or census of document types, properties and conditions, including information on the distribution of document properties and conditions (known as markers), e.g., percentage of paper stock, color paper, bi-tonal text, machine printed and hand-printed text document images in the entire file.

Thus, a file that consists of 1% red color forms would require different technical solutions than one consisting of 55% red color forms. A final pattern analysis will derive images representing marker combinations, e.g., blue paper with blue text and constrained hand-printed entries; blue paper with black and blue text, with unconstrained and constrained hand-printed entries. The method requires that a true random sample of images be pulled from the master file.

6.4 Benchmarking and Verifying Requirements

Running benchmarking tests can provide useful information about the performance, capacity, and capabilities of the systems that will be used to do the scanning. Benchmarking should provide some “real world” metrics that will be specific to the process, staff, and equipment that will be involved in the digitization project.

There will be a wide variety of components involved in the digitization project tasks (e.g. processing, quality assurance, compression, storage, transmission, reproduction, and display). These components might not operate uniformly across a wide range of document with different types and attributes. Exploratory benchmarks can uncover sensitivity needed to restrict component selection.

The exploratory benchmark should run the test set against a variety of hardware and software types. In this way, designers will be able to assess the extent to which such components can handle the test set properties, as well as extract more reliable productivity and accuracy rates.

The benchmark should help to identify optimum scanner resolution levels and throughput performance. It should reveal the frequency that images need to be rescanned because of quality control issues.

The results of such benchmarks should indicate how well commercial hardware and software will handle the different document types within the collection.

6.5 Production and Workload Rates

An additional benefit of running benchmark testing will be the ability to determine reliable workstation production or workload rates. In document digitization conversion systems with a document collection that has a variety of document types, workload throughput rates and workflows can be inconstant.

For instance, assume that a project plans to convert documents to digital images at a steady rate of 10,000 documents per day. However, in practice, the project sees that some days more images are processed, and some days fewer are processed. The variation is related to the variation in the documents to be scanned (if all the documents were essentially the same, the work rate would be much more constant), and the tasks those variations introduce to the work flow. Preparing documents may take longer for fragile physical sources, for example. The quality control comparison may be more time consuming for certain document types, and calibrating the scanner may need to be done more frequently if the document types are varied.

When scanning, for instance, a certain percent of the images will require rescanning after image quality assurance. This will decrease the production rates by some uncertain percent. At the QA station, images will probably undergo several iterations before an acceptable image is attained, thereby increasing the number of images to be handled.

The benchmarking of a good set of documents, which statistically represents the physical collection, will furnish information and statistics about document properties that will enable a more reliable calculation of expected variable rates, as well as a more reliable baseline for determining what resources need to be allocated for the project.

6.6 Document File Statistical Analysis

Once the document sample is assembled, the next step will be to quantify its contents. With any group of documents, there will be a variety of different ways to group or combine the documents. The different groupings will then lead to different statistical summaries representing the individual groups. The combined statistical information for all of the groups should then be representative of the overall composition of the original sample.

The statistics for each group will be based on the metrics determined for each document in the sample. Emergent document groups should be compared with expected groups defined within the project.

The statistics generated will depend on the completeness of the original attribute analysis of each document page in the sample. This first stage should establish a baseline for subsequent and more sophisticated representations used for organizing system requirements for solicitations and system design.

Image based documents can have a wide variety in content; there are almost as many ways to categorize documents as there are different document types. However, focusing on the following main categories will help when trying to establish quantitative statistics for any group of documents:

- Type of document (distributions of forms)
- Type of parameters exhibited (routine or constraining)
- Type of attributes (markers) exhibited (envelope or restrictions)
- Range of attributes (uniformity, complexity)
- Quality or extent of attributes (e.g., few or multiple, a sensitivity measure)

The analysis of types of documents in a collection should provide information about the classes, or document groups that will have to be handled and processed. The different document groups will determine what kinds of scanner settings will be required. This represents the highest, most general category of classification.

Within any document collection there will be similarities of attributes among document types as well as differences. The differences involve structural parameters. The analysis of types of parameters focuses on two types: routine and constraining.

Almost all routine parameters (and associated attributes) are identified as exhibited. They typically include attributes like:

- document size
- stock type
- stock color
- is the document an original or duplicate
- content type

The statistical summary of routine parameters will define the “structural form” of the sample file. This information is useful in defining hardware, software, and workstation functional requirements.

In contrast, the presence or absence and extent (or range of attributes) of constraining parameters define the quality of a document and extent of problem conditions. Thus, if no constraining parameters are exhibited and recorded, the file can be depicted as uncontaminated and faultless from an image processing and quality assurance perspective.

If there are a wide variety of problem conditions, this will indicate problems with human and machine readability, data integrity, productivity, indexing accuracy, system design, and cost controls. This information is useful for determining what types of software operations (e.g. automatic color correction or the use of OCR or ICR software) can be used with the document types.

7 Imaging Processes

7.1 Introduction

A digital image is composed of 2-dimensional, rectangular array of elements called pixels. These pixels have a static value that represents their color. The resultant color array is used to display images on the monitor or instruct a printer on how to render the image to paper. There are a variety of different formats that can be used to store the pixel array which vary the number of bits used to describe the color and use different methods of compressing the data in the array.

Some image processing functions are outlined below. Each of these functions can employ different kinds of hardware and/or software.

The nature of the material being scanned and the requirements of the project will determine the appropriate quality control scanning procedures that should be implemented. Determination of an image's quality should be based on the research needs of users (and potential users), the types of uses that might be made of that material, as well as the artificial nature of the material itself. There is no single set of image quality parameters that should be applied to all documents that will be scanned.

7.2 Pre-Scanning Image Enhancement and/or Restoration

This category of tasks covers preparation of the physical image. There may be some techniques that should be applied to the physical document to improve the visual quality of an image. These techniques are normally used to restore the image or image surface. This portion of the image processing may be done well in advance from the actual scanning. Digitization projects should consider training scanners to look for physical conditions that may require restoration, if restoration is an option.

Other tasks include grouping documents into like categories. Document grouping simplifies the task of calibrating of scanning equipment because the scanners only need to be adjusted before each new group of documents is scanned, not before each individual document.

Adjusting the scanner calibration (adjusting contrast/brightness enhancement, edge detection, cropping, alter-tonal or color settings, selection of correct resolution, reducing noise, etc.) will improve the readability/legibility of the digital image. See Section 9: Scanner Calibration and Quality Control for more information.

All equipment surfaces must be clean and dry before being used with records. Cleaning and equipment maintenance activities (e.g. replacing toner cartridges) should not take place in the same area as normal scanning, in order to prevent spills or other accidents that could damage the originals. Aerosols or ammonia-containing cleaning solutions should not be used. A 50% water and 50% isopropyl alcohol solution is recommended [2] for cleaning. Cleaning supplies should be stored away from the document processing area.

7.3 Document Preparation

Delicate originals should be placed in non-glare transparencies so they are not damaged in handling or by the operation of the scanner. The initial document survey should have identified documents that will need to be handled in this way, as well as providing samples for scanner operators so that they can recognize what constitutes a "delicate" document and how it should be placed in the transparency.

Image placement is also important. Correctly aligning documents and ensuring that automated feeders are working properly will help minimize jagged breaks in horizontal and vertical lines.

Oversized documents, which will be scanned multiple times to create a single large image, should be carefully positioned to ensure that no area is missed in the scanning process.

7.4 Scanning

Scanning is the actual process of creating a digital image from a physical document using a combination of scanning hardware and software. The scanning hardware uses reflected light to create a digital impression of the document, which is then converted into a specific file format by the scanning software. The composition of the physical image and the specifics of the scanner hardware determine what data will be stored in the digital image.

This task is the core of a digitization project. The correct scanner must be used and correctly configured. Documents must be accurately placed on the scanner-bed, or placed in front of the camera lens properly in order to create a proper digital image. Because of the optical and organic nature of this process and the mechanical nature of the hardware involved, this process is subject to the most variance.

The nature of the material being scanned and the requirements of the project will determine the appropriate quality control scanning procedures that should be implemented. Determination of an image's quality should be based on the research needs of users (and potential users), the types of uses that might be made of that material, as well as the artificial nature of the material itself. There is no single set of image quality parameters that should be applied to all documents that will be scanned. The best situation is one where the source materials and project goals dictate the image quality settings and the hardware and software that will be used.

Because scanning is subject to more variance than other areas of image processing and because this stage represents the transition from the physical to digital world, this is the area that should receive the most attention to quality control.

7.5 Image Enhancement

Image enhancement is meant to improve the usability of a digital image. Enhanced images are created from master images; they should not replace them. Image enhancement processes can be used to improve image capture but their use raises concerns about fidelity and authenticity. Use of these processes can also dramatically increase the cost of conversion. Image processing filters -- mathematical formulas that change the appearance of digital images -- can be applied to improve the appearance of images and to assist with resizing images.

Image enhancement can involve a combination of manual and automatic procedure to deskew, sharpen, and remove noise from digital images. Commonly, sharpening filters are used to enhance the appearance of digital image files. The need for sharpening is inversely proportional to the resolution of the digital image: lower resolution or smaller digital images tend to need more sharpening, and higher resolution or larger digital images tend to need less sharpening. A common sharpening filter is unsharp mask. This term comes from the graphic arts industry practice of using a reverse toned mask that is slightly out of focus to increase the visual sharpness of images. It is possible to over-sharpen an image: Over-sharpening with an unsharp mask filter will create light halos around sharp edges within images.

Another filter commonly used when resizing images is the blur filter. Slightly blurring an image creates additional shading along sharply defined edges in an image, which can allow the interpolation software to do a better job when the image is resized. Most images have to be sharpened after resizing, whether or not a blur filter is applied.

Just as with interpolation algorithms, some image-processing filter algorithms will do a better job in terms of image quality than other algorithms, while others might work faster. Again, generally the filters in more expensive image processing software will tend to do a better job with image quality compared to the filters in less expensive software.

A common image-processing tool is the histogram, found in most image processing software packages. The histogram is a graphic representation of the distribution of gray shades in an image. The height of each vertical

line is proportional to the number of pixels that are of that shade -- the taller the line the more pixels of that shade. Also, the histogram can give indications of certain types of image defects, such as loss of tones in the shadows (dark values or shades) or the highlights (light values or shades) of an image.

Thresholding is a technique used in image processing to convert gray shades to either black or white. All shades lower than a selected value are rendered as white and all shades higher are rendered as black. Depending on the value selected for the threshold, the representation of the same image can be altered dramatically. Most 1-bit scanners actually sample at 8 bits, but then a threshold value is used to convert the 8-bit image to a 1-bit image. In cases of thermofax, verifax, or carbon copy processes where the paper ages as it darkens and the type fades, it is very difficult to reproduce the image with a 1-bit scan regardless of the threshold level. At lower threshold values the characters appear incomplete. As the threshold value is increased, the characters will quickly fill in (*e.g.*, the letter "o" becomes a very large dot) and only the context within the word or sentence provides an idea of the character. Further increasing the level of the threshold will cause pixels representing shading in the background to turn black, an effect that is known as speckle. There are software programs designed to work with 1-bit scanning designed to despeckle an image. The software tries to remove extraneous black pixels in the image. Unfortunately, this doesn't always work the way you want. Parameters for despeckling can be adjusted, based on the size of the speckle you want to remove, but as the size of the speckle to be removed is increased, it will start removing periods, dots of "i"s, and other necessary punctuation.

When using low bit depth images, it is possible to simulate a greater number of shades with fewer shades. This process is known as dithering. The key is to redistribute pixels according to a mathematical formula to produce synthetic shades of gray based on the arrangement of these pixels and the way the eye perceives them. There are different formulas for dithering, and some work better than others. If an 8-bit grayscale image is converted to a 3-bit image without dithering, broad areas of similar shades will be rendered as a single shade. In digital images, this effect is sometimes referred to as banding, particularly when it appears across broad shade gradients, such as skies in photographs. When a 24-bit color image is converted to an 8-bit color image, the 8-bit file can be dithered. Dithering and an adaptive grayscale palette can be used to provide a very accurate rendition of an image with bit depths as low as 4-bits or 16 shades.

Deskewing can help remove stair-stepping (or "jaggies"), which engineering and line drawing are especially susceptible to.

Whenever enhancements are made to an image, they should be noted in the metadata associated with the file. The file should be stored (and indexed) as a derived image of the master. **It should never replace the original digital master.**

7.6 Image Compression

Recall that the digital image is composed of a fixed number of pixels aligned in an array and that each pixel uses a certain number of bits to represent a color. The total size of a file in its raw form is equal to the product of the height, width, and number of bits per pixel.

Image compression is the technique of reducing the size of digitized documents when they are stored on disk. Reducing the image size means that more images can be stored in the same amount of space and that files can be transferred more quickly over the network (an important consideration for web-based access).

The amount of compression gained for any specific image depends on the type of compression algorithm used and the composition of the image. It is not uncommon to see compression rates of 10:1 for text based forms and 2:1 for film or continuous tone photos.

There are many different types of compression available, all with different efficiency ratios and algorithms. Certain compression methods are more suitable for certain kinds of data. Some algorithms are "lossy", meaning that some data is lost during the compress, and some are "lossless", meaning that all of the data is recovered when uncompressed. In general, "lossy" algorithms provide a higher degree of compression (at the cost of data loss).

The following figure [29] shows the original image stored in a lossless format on the left (GIF) and the same image then stored in a lossy format (JPEG) on the right.



Figure 7-1 An image in lossless format (left) and lossy format (right)

The most common image compression formats are outlined in Section 14: Image Formats.

Lossy compression algorithms gain extra compression efficiency by reducing the amount of information in an image. For example, a lossy compression algorithm may combine like colors (e.g. blood red, dark red, maroon, etc.) into a single color, thereby reducing the total number of colors in an image and hence the total number of bits needed to represent all the colors of the image.

The transition from an uncompressed image (rasterized image) to a compressed one is usually handled automatically by scanner software when it stores the image on disk. The format selected for image storage determines the compression algorithm that will be used.

Care should be taken in selecting image formats to use. In almost all cases a lossless compression format should be used for master copies, and a combination of lossy, compression and/or reduction in resolution can be used to create view or access images.

7.7 Image Analysis

Image analysis is the technique of measuring and classifying information attributes within an image, such as indicators of shape, descriptions of outlines, brightness and color. This includes both automatic and manual classification. Automatic analysis encompasses OCR and ICR⁷ as well as automated software analysis.

7.8 Indexing

Indexing creates records in a central “catalog” database that uniquely identifies each image and where that image is stored. Good indexing will also include links to the metadata associated with the image file and other related images (i.e. thumbnails, access images, enhanced images, etc.). Indexing bridges the gap between storage and retrieval.

Image indexing is most effective on highly referenced collections where a short retrieval time is important or where there are multiple users accessing the same records.

An index entry can consist of a person’s name, title of a folder, or just a date and page number. Indexing can occur through manual data entry on a workstation (with the document image located on a segment of the monitor screen) and/or through automated optical character recognition.

⁷ Optical Character Recognition and Intelligent Character Recognition. OCR is a software technique of determining what text is present in a digital image by analyzing the pixels of the image. OCR is generally used with images of machine printed characters because it relies on a font library. ICR is more flexible and can be used to read handwriting and other irregular text, but requires an operator to “teach” the software so that it can learn from its mistakes.

Good indexing requires an explicit structure and entry policy to ensure that images can be uniquely identified. Strict quality control over the indexing of data also ensures that data images can be located accurately for future use. The developed QCMBT tool (see Appendix D) is designed to assist in this task.

If an image is created from multiple smaller images, such as is the case when a document is too large to be scanned in its entirety, proper indexing will ensure that all of the smaller image files are associated with the larger main image, and can be accessed by users if needed.

In general, if the data associated with the indexing can be self-describing, it will have a higher degree of integrity over the long term. Metadata refers to self-describing information, and is outlined briefly in Section 12: Metadata.

7.9 Distribution and Transmission

The primary constraint for image distribution is the size of the image itself. Networks have a maximum speed at which they can move data; a certain number of bytes per second. Larger images will, therefore, take longer to move through a network. If network speeds are slow, compared to the size of the image, the retrieval will be slow. Network speed may be a major consideration if images are to be accessed over a dial-up connection.

Creating derivative images (access images and thumbnails) specifically for remote access can reduce the burden on the network. However, selecting a file format with a good compression algorithm and a tolerable lossy factor can also significantly reduce file size and improve access over the network.

Scanning resolution differentials (e.g., 200 vs. 300 DPI) affect transmission speeds because higher DPI images will be larger (in terms of data size, not viewable image size) than a low DPI image made from the same physical document.

7.10 Quality Control

Quality control ensures all processes meet engineering standards, that all functional and performance specifications are achieved and maintained, and that all resulting images achieve fidelity, legibility and readability. In short, the goal of quality control is to make sure that the images created during scanning are good enough to meet the needs of the project.

There are two primary quality control tasks. First, the scanner should be properly calibrated. A combination of standard test targets and sample documents will provide metrics that can be used to determine how accurately images are captured by the scanner software. In some cases automated software tools are available to do numeric analysis on the images created during testing and verify that they meet required quality levels.

For more information on scanner calibration and the use of test targets see section 9: Scanner Calibration and Quality Control and section 10: Sample Standard Tests.

The second quality control task is more subjective. One or more quality control personnel should visually inspect the digital image and compare it with the original. Check that the entire image has been captured (i.e. not cropped) including any captions or titles. Are pages missing or out of sequence? Is the image skewed? Does the image have the correct file name? A second set of checks is more complex to assess, and includes detail reproduction, tone reproduction, color reproduction, and color accuracy. For images of textual material, line drawings, etchings, plans and other objects with distinct line-based features, detail reproduction is the key to image quality. When benchmarking, a resolution target or the smallest resolvable detail should be used. This provides a comparison point for examining legibility, completeness, sharpness, contrast, serifs and uniformity, paying particular attention to individual strokes and dense cross hatchings. The visual inspection should detect visual aberrations as well as any interference patterns (i.e. moiré patterns) or other visual effects that rob the image of necessary information.

For grayscale and color images the bit depth and dynamic range are as important as resolution in assessing image quality. Assessing color and tone reproduction can be highly subjective, particularly if fidelity is desired, but features to look out for include the presence of details in shadows and highlights (an indication of a good dynamic range), and a smooth transition in tones, particularly on skin and sky (a blotchy or pixellated effect is an indication of insufficient bit-depth). Compare color, contrast and brightness to the original or to a color chart, paying particular attention if digitizing from negatives.

The digital master may not be a perfect representation of the physical object, but it will be up to the quality control inspector to determine if it is of a suitable quality for use in the project.

7.11 User View

The user view process consists of image and data access and retrieval; including access software (association with physical data base management for imagery and image indices) and workstation monitor screen resolution, printing, faxes, and related controls.

This is also the stage where compressed images are uncompressed. Decompression takes place at the software layer. Some software includes support for uncompressing popular file formats (e.g. web browsers automatically uncompress JPEG files for display). If a proprietary storage format is used, the software employed by the end users will also have to be able to uncompress (and display) data from the proprietary format.

8 Scanner Equipment Considerations

The upper limit of quality for digital images created by scanning is primarily dependent on two factors: the scanning hardware and the scanner configuration. This section briefly outlines some points to consider when acquiring scanning hardware, including general recommendations encountered as part of the general quality control survey. A full examination of scanning hardware is outside of the scope of this project.

Quality control procedures for scanner calibration are discussed in Section 9: Scanner Calibration and Quality Control.

8.1 Scanner Selection

You should be aware that not all scanners work the same; and, possibly, that not all scanners will be able to process all the types of the source document images in your files. In the same way, not all data capture OCR software packages will work the same or yield the same results and error rates on a given original source document's image.

You can expect varying levels of acceptable and quality results of document image processing, due to many contributing factors. It is even possible that an original source documents could be processed differently by different, supposedly compatible, technologies.

The key to consistency and stability is especially the connectivity between imaging technology and the source document images.

8.2 Flatbed Scanners

Flatbed scanners are one of the most popular scanners used and are suitable for scanning papers, flat photographs, and other printed materials. Flatbeds can be purchased with an optional attachment called a transparent media adapter, which allows you to scan directly from slides or negatives. However, transparency adapters do not always produce as high a quality of image as a slide or film scanner.

There are also flatbed scanners that handle originals that are 12" x 17", and some flatbed scanners can accommodate even larger sizes [13], although they tend to take up considerable space and produce enormous file sizes.

There always exists the possibility that documents will be larger than the scanner bed, no matter how large the scanner. In these cases, multiple smaller scans can be made from the original, then pieced together into an original document. Flatbeds represent the second best choice, after high quality digital cameras, for dealing with oversized documents, since they have the fewest moving external parts or constraints that could damage the portions of the document that extend beyond the scanning area.

Automatic feeders, which can improve work rates, may be available. But automatic feeder devices can also damage documents that need to be handled delicately, so they should only be used for documents that will not suffer from the automated feed process.

Flatbed scanners use a linear CCD⁸ array, made up from a long line of CCD elements in a row. The CCDs themselves can only detect the presence or absence of light. To enable the scanner to capture color, they must either make three passes with a Red, Green or Blue filter in front of the CCD or have 3 lines of CCD each with

⁸ Charge-Coupled Device. A semiconductor technology used to build light-sensitive electronic devices such as cameras and image scanners. Such devices may detect either color or black-and-white. Each CCD chip consists of an array of light-sensitive photocells. The photocell is sensitized by giving it an electrical charge prior to exposure.

either a red, green or blue filter on top. Recently, CCD based scanners are much cheaper to produce than PMT⁹-based scanners and also tend to be much easier to operate. They range from very cheap and low-end consumer devices up to professional quality devices with costs comparable with cheaper drum scanners (in the 10's of thousands).

For the digitization project one of the main advantages of the flatbed scanner is that they are generally much simpler to use, allowing unskilled operators to effectively use them without many weeks of extensive training.

8.3 Slide Scanners

A slide or a film scanner [13] should be considered for scanning predominantly transparent materials that are smaller than 4 x 5. There are also some slide/film scanners that can handle larger transparent formats. Scanners that combine flatbed scanner capabilities and 35mm slide capabilities are also on the market.

Some slide scanners can deliver a better dynamic range than flatbeds; however, the resolution may not be sufficient to create digital masters or meet the resolution requirements of some users.

8.4 Digital Cameras

Collections with a large number of oversized documents may want to consider purchasing a high-end digital camera. These digital cameras should be able to capture oversized materials. They can be configured to work much like a copy stand setup. This is very important to institutions with bound volumes of oversized documents, three-dimensional items, and special items. Unbinding, size reduction by cutting, and anything which results in the defacing of these resources in order to facilitate either microfilming or scanning at this time should be given careful consideration.

Some digital archivers [13] feel that commercially available, hand-held digital cameras are not suitable for archival scanning, excepting the high-end digital cameras¹⁰ used by several larger institutions and imaging vendors. High-end digital cameras don't have scanning limitations when it comes to size and shape, and can scan at an extremely high resolution (up to 15,000 pixels across the long dimension). However, they do require certain lighting requirements and a high level of operator skill.

Digital cameras can also take longer to set up than traditional scanners. However, when working with oversized documents, using a digital camera may provide better work rates than using a scanner with a smaller image area to create a single large image file from multiple smaller images.

Digital cameras can present great potential for scanning oversized materials, media in all formats, and bound materials with the aid of a book cradle. They also pose a lower risk to fragile materials by allowing face-up, contact-free scanning.

However, at this time, the only digital cameras that meet most guidelines are extremely expensive. Fortunately, their capabilities and prices are rapidly improving and may become the ideal scanning device in the future.

8.5 Drum Scanners

Drum scanners are designed for the graphic arts community and, as such, provide an extremely high level of resolution. Drum scanners can scan transparent as well as reflective media, in grayscale and color.

However, drum scanners are not recommended for documents that are fragile or brittle in any way, as drum scanners can cause a great deal of stress to the document. The original document must be flexible, as it will have

⁹ Photomultiplier Tube. PMT are very sensitive to small changes in energy (i.e. light), which are used to distinguish between different light intensities at each pixel.

¹⁰ Kontron, Zeutschel, Leica are some makers of high-end systems that have been used historically to create digital archives

to be secured (sometime by taping or clamping) to the rotating cylinder, which may damage or discolor the original document.

Drum scanners use photo-multiplier tubes (PMT) to produce very high quality results. They typically have a Density Range of 3.4 – 4.0 with a 'dMax' at the top of that range. They can offer an optical resolution of up to 8000 samples per inch (spi). Due to their complexity they require skilled operation. The documents is 'mounted' in the scanner on a transparent acrylic cylinder (drum) and then spun at high speeds around the photo-multipliers within the cylinder. Mounting transparencies on the drum is a slow and skilled operation and it is normal to have at least two drums in use so that one can be mounted whilst the other is being scanned.

Although the quality from these scanners is exemplary, they tend to be slow and may not be able to provide the level of productivity required from many digitization projects. These trade-offs should be considered in any consideration of their use.

Drum scanners typically cost many tens of thousands of dollars although there have recently been a few desktop drum scanners introduced at a more affordable rate.

8.6 Hand-Held Scanners

Hand-held scanners are scanners that are moved or rubbed across the face of documents. They are generally not recommended, as they tend to introduce additional artifacts into the digital image.

Additionally, hand-held scanners can cause damage to fragile documents. According to NARA preservation specifications,

"Equipment that could potentially damage documents will not be approved. No part of the equipment may come in contact with records in a manner that causes friction, abrasion, or that otherwise crushes or damages records...." [2]

If a hand-held scanner must be used with fragile documents, it is recommended that the documents to be scanned must be protected by polyester sleeves. The sleeves should be larger (at least one inch larger on each side than the document is a good guideline) than the document to be scanned to protect the edges of the document from inadvertent damage during the scanning process. For preservation reasons, the scanner should never come into direct contact with the document.

8.7 Scanner software

There are two types of software that you will need for most digital imaging projects. The first is the scanning software that comes with the scanner. The second type of software is the image editing software, normally applied to the image after it has been scanned. Some software, such as Adobe Photoshop®, can serve as both the scanning software and the image editing software.

The scanning software is usually limited in its functionality. You should choose scanning software that is at least capable of saving image files into standard formats such as TIFF, JPG, GIF, etc. This functionality will help production and also ensure a wide range of image delivery options. Software that converts image files from one format to another may also be useful.

To produce images of acceptable quality, it is important to invest in image editing software, which is normally used for "cleaning up" an image (removing dust spots, for example) and for correction (adjusting the level of brightness and contrast, for example). Image editing software should come with the capability to crop, de-skew, and rotate; adjust brightness and contrast levels; sharpen (if needed); zoom in and out; accommodate different file formats; provide controls for gamma, black and white, and color (RGB); provide a histogram and look-up table; support compression types; and possess the capability for the user to create and save customized settings, among other functions.

The choice of image editing software is based on the level of image manipulation desired for your project and the level of expertise of staff. Some image editing software, such as Adobe Photoshop®, is very advanced, and may require some time and training to learn. Other software is more basic and allows for only limited operations, such as cropping and rotating, and is not difficult to master. Consider the range of operations you will normally need to perform. The cost of this software can range from free (freeware) to several hundreds of dollars. When considering cost, think about not only the cost of the product, but also how easy it is to use—and factor in additional costs for training, accordingly.

In addition to considering the capability and usability of image editing software, make sure that the project's current technology can support the software. Computer systems should have an appropriate amount of memory, hard drive space, processor power, and display capabilities (a 24-bit color display card is recommended for image editing work).

The amount of image editing performed (if any is to be done) on the images should be defined in the project goals, possibly decided in consultation with the collection curator or an archivist or librarian who is knowledgeable about the materials being scanned. Some digitizing projects aim to create a "pleasing image" that may require a great deal of editing. Other projects may be more concerned with the fidelity of the digital image to the original (this may be important to scholars), and may require very minimal editing.

At the very least, most projects will want to create a thumbnail image and an access image. The creation of these images will be done through the use of image editing software. When evaluating acquiring new image processing software, determine how easy it will be for project staff to create the derived images. If the derived image are to be of different sizes, enhanced, or cropped, consider also how easily these tasks can be achieved using the image software.

Editing should only occur on derived images, never on the master image itself. The specifics (e.g. how and why the image was altered) of the changes should be stored in the metadata. The indexing should link the master and edited file.

The software processing needs of the project will be different depending on if the goal is to match the digital image as closely as possible to the original or if it is more concerned with capturing the photographer's/creator's intent when editing the digital image. Or the project goals may be to reconstruct the appearance of the original as it would have existed when it was first created, which may require digitally reconstructing deteriorated originals.

8.8 Scan Rates

Not all scanners take the same amount of time to scan the same image at the same resolution. If high production levels are important, it will be necessary to look at the time it takes for both preview and full scan images of materials similar to what you intend to scan.

In general, flatbed/slide scanners accommodate a higher production rate than digital cameras, but they also are limiting in the size and type of media formats they are able to scan. Some flatbed scanners are available with automatic document feeders that can improve work rates, but which may not be suitable for all document types.

Certain models of scanners may be better suited to the project. Consider scan rates and approaches. It may be cost effective to acquire several types of scanners to meet the specific needs of the project based on the distribution of document types within the collection.

8.9 Scanner Resolution

The resolution at which an image is scanned is one of the factors that will determine the quality of the digital image that is produced.

Resolution is often expressed as an array: the number of pixels across both dimensions of an image (or more simply as 3000 pixels across the long side), as DPI (dots per inch), or as PPI (pixels per inch).

Higher DPI settings will generally yield a better digital image, because they place more pixels (therefore, information) in an inch than do the lower DPI settings. However, the higher the DPI, the larger the file size will be. Project planners should take into account server or computer storage capacity when determining resolution settings, and balance that against the goals of the project.

Scanning at a high resolution is recommended to generate "archival" images, or make prints of the digital image on a good printer. There is a threshold to resolution, however. After a certain point, increasing resolution will not yield a better image.

Image quality tends to level out as scanner resolution increases. The following figure [29] shows the conceptual gains in quality when moving from low resolution scanners to higher resolution scanners, and that the overall image quality levels out as high resolution scanners are used.

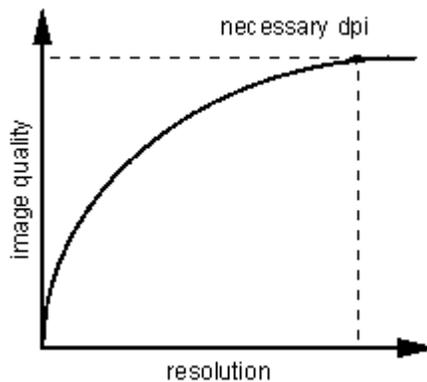


Figure 8-1 Relationship between Image Quality and Scanner Resolution

There are two different types of resolution: optical and interpolated.

OPTICAL RESOLUTION

Optical resolution is the inherent resolution of the scanner, and is usually expressed as a pixel array (i.e., 1000 x 2000). The first number refers to the limit of the CCD array on the scanner (the short dimension), and the second number refers to a number determined by the movement of the CCD array across the long dimension of the scanner.

A single pixel in a digital image can only have one color. However, in the original physical document, that same area may have more than one color value. During digitization, the average color value for that area will be used (based on the reflected light energy).

As the pixel area shrinks, the risk of having multiple colors blended into one decreases. A higher optical resolution will result in a smaller individual pixel area.

INTERPOLATED RESOLUTION

Interpolated resolution is calculated by software from a lower resolution image file. This is often performed during or after scanning.

A better digital image will be created by a higher optical resolution scanner than will be created by interpolated resolution.

The specifications for the resolution at which you scan should represent actual optical resolution rather than values achieved by interpolation.

8.10 Noise

All electronic devices [13] suffer from "noise," which often appears on scans as blotchy or matte-like areas in the dark shadow parts of an image when these areas are lightened or have their contrast range increased. Noise can obscure details in the shadows. Higher quality scanners, with higher bit depths, will give better results, as they tend to use higher quality (lower noise) components.

8.11 Light Sources

Some light sources [2] are capable of raising the surface temperature of the record being scanned. If this temperature is raised high enough, it may damage the original document, particularly if the original is a fragile or deteriorating.

Light source heat specifications should be included in any RFP when acquiring new equipment.

8.12 Scanner Suggestions Based on Material Type

SINGLE LEAF, REGULAR SIZE, FLAT MATERIALS

A flatbed scanner is probably the best choice, as it provides the most stable platform for scanning. Non-brittle documents can be scanned using a sheet fed scanner, which will improve scan times, but can be more subject to alignment issues than hand positioning.

High-resolution digital cameras can also be used effectively, but low-end digital cameras will generally produce lower resolution images that are not suitable for use as digital masters.

SINGLE LEAF, OVERSIZED, FLAT MATERIALS

An oversized flatbed scanner is probably the best choice, as it provides the most stable platform for scanning. Non-brittle documents can be scanned using a sheet fed scanner, which will improve scan times, but (as mentioned previously) can be more subject to alignment issues than hand positioning.

If no oversized scanner is available, a low-end digital camera can be used to produce non-master images. This will generally produce a lower resolution image than using a scanner, but may be the only physically feasible option. The image in this case may only be suitable for use as a thumbnail image.

High-end digital cameras, on the other hand, present an excellent alternative, because they can accommodate virtually any size of document.

If the documents can be manipulated without causing damage, a smaller scanner can be used to scan portions of the document. The smaller images can then be pieced together using image editing software to create a single larger image. Resources permitting, a full image should also be taken and associated with the sub-images, even if the full image is taken at a lower resolution (such as would be the case if the full image were taken with a low end digital camera).

BOUND MATERIALS

Bound materials present two main problems in that they may not lay flat when opened, and that opening them fully may damage the binding. Bound materials limited by these problems should be scanned using either a digital camera with a book cradle or a right angle, prism, or overhead flatbed scanner.

TRANSPARENT MEDIA

Transparent media should be scanned using either a slide or film scanner. Multi-format flatbed scanners may also be suitable.

To some extent, high-end digital cameras can be used, but because some transparent media (for example, microfiche) can have a relatively large amount of visual data stored in a relatively small amount of space. Digital cameras may not be able to capture this level of detail.

PHOTOGRAPHIC IMAGES

Color photographs using Ektachrome, Ansco and some other early types of color film during the 1940-1982 time period are among the most fragile photographic items [9] in most collections.

Nearly all of them are experiencing dye instability that cannot be reversed.

These photographs can be stabilized, but in many cases the deterioration is already significant. Scanning these photographs or negatives, repairing the image in photo software, and printing using the best possible archival systems (printer and paper) offers one of the most affordable current means of preservation.

Ideally, the prints would each be photographed on more stable current print paper. Realistically, most institutions cannot afford to do this. Scanning in very high resolution files can be considered an archival option for these images and may be the only instance where the digital file is more stable than the deteriorating original.

These considerations are related to film media itself, not the method by which the photo was taken. As such, they apply to aerial photos and photos taken with more traditional film cameras.

9 Scanner Calibration and Quality Control

A large portion of the quality assurance procedure relies on the subjective expertise of the operators. Properly calibrating scanner systems will increase image accuracy, and therefore objectivity, in the scanning process, reducing the subjectivity of image evaluation.

9.1 Image Quality

The definition of image “quality” depends on the project’s planning choices and implementation. Project designers need to consider what standard practices they will follow for input resolution and bit depth. Projects may need to consider what image layouts will be used, if cropping should be allowed, and determine what image capture metrics (including color management) are appropriate.

Benchmarking, based on a sampling of a variety of documents in the collection, for different types of source material can provide a quantitative measure of quality. Those measurements can then be used to determine what configuration makes sense so that the images produced will meet the quality goals of the project. They can also be used to compare the results of day-to-day scanning quality control.

By maximizing the image quality of the digital master files, managers can ensure the on-going value of their efforts, and ease the process of derivative file production.

9.2 Quality Control Advantages

Projects that use benchmarking to specify a minimum level of acceptable quality for good digital images can improve the value of the digital images to the end user and minimize the risk of losing valuable information.

Creating good digital images will improve their long-term viability and reduce the risk of needing to re-create the images at some future date, even as production techniques improve.

If a project is able to produce good digital images, the users of those images will develop confidence in them because they will have a minimum level of well-known and consistent properties, and will support a variety of known uses.

If problems occur during scanning, but the documents are not examined until they are needed, the documents will need to be rescanned. The user will also have to wait for the data. Sometimes the original documents are discarded or archived after scanning, making it difficult or impossible to acquire the original for rescanning. Therefore, image quality should be verified while the original document is still available.

9.3 Quality Control Considerations for Scanner Calibration

Different document types will generally require different scanner settings, calibrated to the appropriate configuration that best suits the target. In order to control the quality of digital imaging when the scanning hardware configuration is often changed, it is necessary for the operator to easily be able to verify that the scanner has been configured properly.

There are three main areas of quality control concern when it comes to the operation of the actual scanner.

First, the operator must understand the capabilities of the scanner, i.e. what it will do and what it will not do. This area of quality control is related to user training. Project planners should anticipate training requirements based on establishing a basic level of instruction in scanner use, and then providing specific training related to individual scanners. The project should also allow for additional training if new equipment is procured.

Then, before the actual scanning process begins, the operator must be able to verify that the scanning equipment is setup and functioning properly. The operator will also verify that the scanner was still functioning properly after all the documents were scanned.

The operator should be able to set up criteria for quality control based on specific needs of the documents being scanned. The operator should be able to establish and test new quality control references as new document groups or needs are encountered.

The use of test targets and quality control references can make it much easier for operators to confirm that the scanner is operating at an acceptable quality level and that the configuration is appropriate to the document type.

Finally, each of the images in the sample will be inspected. Recognizing the good and bad artifacts or properties can be very subjective and therefore difficult for new operators. For training purposes, illustrations of real images with properties that are problematic, plus other descriptive recognition criteria should be made available.

Some digital archivers recommend that the visual inspection process be done by someone other than the person who did the initial scanning, under the assumption that a second person will be more objective. However, delaying the review of images until after the scanning can introduce a long delay between “bad” images are produce and when the problem is detected. All of the images produced during that time frame will need to be examined; and possible re-acquired and rescanned, causing a long delay in digitization.

A better approach is to do the visual inspection at the time images are created, with an optional “peer review” of some or all of the images before they are released for use.

9.4 Scanner and Monitor Calibration

Most scanners and image editing software provide functions for calibration of the scanner and/or monitor (including monitor brightness, contrast, and control of gamma settings). Scanner software is often used to match the tonal scale of what is being scanned, which may include black and white or color calibration. In general, a check of scanner calibration should occur every time you scan a new media format or scan a new media size.

Scanners should be adjusted to provide the best possible results based on the type of document scanned.

Computer monitors can misrepresent the scanned image if not properly calibrated. Image characteristics, such as moiré, wavy lines, dark or light spots, inaccurate resolution, etc. may be introduced if the monitor is not calibrated.

Suggested Monitor Settings:

- Set to 24 millions of colors
- Set Gamma at 2.2 (1.8 for Macintosh)
- Color temperature at 6500° K
- Calibrate to RGB (Standard Default Color Space for Internet)

Monitors should be calibrated regularly. There is specific software you can purchase to calibrate your monitor.

Monitor calibration guidelines will help users adjust the brightness and contrast of their computer monitor so that digital images will look their best. If computer monitors are adjusted to a target, the digital images (if scanned properly) should provide a reasonably accurate depiction of the originals when viewed on the "average" computer monitor.

Calibrating the monitor is essential for the visual inspection of digital images. If the presentation of a digital image isn't accurate, the visual comparison won't be either.

9.5 Quality Control Reference Targets and Color Bars

Targets are used to verify the tone and color reproduction of the materials you are scanning and are also used to measure system resolution (targets are about the scanning system and the accuracy of the system to reproduce correct tonal values, not about the materials that you are scanning).

Tone reproduction refers to the degree to which a digital image conveys the luminance ranges of the original. The ideal in tone reproduction is to match the various brightness levels in the original with the brightness levels in the digital reproduction. This is not often achieved, since the digital image is different from the original, and viewing conditions are also different. What can be achieved, however, is an acceptable subjective tone reproduction that can give an approximation of the luminance ranges in the original. Targets provide a means of controlling tone reproduction.

Targets are a way of predicting image quality, and help ensure that the scanning system you are using is producing the best quality image it can and is operating at a consistent level of quality over time.

See Section 10: Sample Standard Tests for target samples.

Note: the original targets should be requested from the supplier, since targets made from photocopies or laser printers will suffer from the same potential image degradation that is trying to be measured on the scanner.

Different targets for prints and transparencies exist. Targets must consist of the same material as the media being scanned (paper, film, etc.). Targets usually contain patches of color, black and white, or shades of gray for verifying tone reproduction. Some examples of targets include the Kodak Color Separation Guide, Grayscale Control Bar, AIIM Scanner Test Chart, IEEE Standard Facsimile Test Chart, and the RIT Alphanumeric Resolution Test Object target. To ensure color fidelity from scanner to monitor, the use of color targets and proper calibration of the monitor is recommended. Some color targets are the Macbeth Color Checker Rendition Chart and the PostScript IT8 Color Output Target.

Some digitization projects are also scanning a color bar along with the original, to be included in the final digital image, to aid users in verifying accuracy in color reproduction. Color bars use colors of a known “value”. Both the color itself, and the value used are included in the color bar for reference.

It is extremely important to establish a quality reference for what defines a “good image”. Using quality references allows non-technical users to judge whether or not an image is of acceptable quality. When making quality reference materials, both the original hard copy and the digital copy must be preserved. The scanner configuration, settings, and any other important information should be preserved along with the hard copy.

If multiple document types are to be scanned, multiple scanners are to be used, or multiple configurations are to be employed, a quality reference should be created for each possibility.

A well-calibrated grayscale target and standardized color target should be used for measuring and adjusting the capture metric of a scanner or digital camera.

For capturing images from reflective media, some digital archivist [12] recommend that a standard target consisting of grayscale, centimeter scale (useful for users to make sure that they are printing or displaying an image at the right size), and standard color patches be included along one edge of every image captured, to provide an internal reference within the image for linear scale and capture metric information.

Alternately a single copy of the targets could be created for each document batch. These samples could then be linked to all of the digital masters in the group through metadata and indexing. This preserves the original document, reduces the imaging overhead of adding the targets to all scans, but at the same time provides calibration references to the user, should they be needed.

A document test set therefore represents a source of mandatory requirements, a vital input for system planning, exploratory benchmarking, and development of system concepts and specifications for solicitation preparation. It provides a baseline for setting the capabilities and boundaries of potential hardware and software functions that can handle or process the disclosed attributes and attribute counts. It also provides the baseline for estimating workstation production rates.

Measuring the accuracy of visual information in digital form implies the existence of a capture metric (i.e., the rules that give meaning to the numerical data in the digital image file).

For example, the visual meaning of the pixel data Red=246, Green=238, Blue=80 will be a shade of yellow, which can be defined in terms of visual measurements. Most capture devices capture in RGB using software based on the video standards defined in international agreements.

Imaging projects should adopt standard target values for color metrics, so that the project image files are captured uniformly.

9.6 When to Run Quality Reference Tests

Frequent testing is recommended [22], because it reduces the loss of time and data associated with needing to rescan documents. If it turns out that images need to be rescanned, the longer the delay between when images were created and when it was determined that they needed to be rescanned, the larger the number of documents that need to be reprocessed.

As this delay grows, so does the risk that the original documents might not be readily available (or available at all). There is also the potential risk that arises from not being able to meet the needs of the user requesting the information in the document.

Quality tests should be run at the beginning and end of work shift changes, when the scanner is changed, and at the beginning and end of scanning for document groups. This will provide a good balance between re-scanning risk and the overhead of running the tests.

WORK SHIFT CHANGE

If the work shift is to change in the middle of a batch of document processing, then a “post-test” should be performed at the end of the shift. The next shift can then perform a new “pre-test”. This helps reduce variance between operators. If there is a long break between scanning sessions, it is possible that the scanner may have changed between uses. Testing at the start and end of every work session can eliminate this risk.

RECALIBRATION

The tests should be re-run immediately after any adjustments have been made to the scanner or any recalibration has been done. If a technician was used to make changes to the scanner, this will also provide a method of verifying that the recalibration was successful. Running the test immediately after changes have been made to the scanner will allow operators to determine if the system still needs to be tuned without having to call the technician back a second time.

Any time the scanner calibration is changed, quality assessments should be performed on targets and the results logged. The scanner settings should be noted along with the quality reference materials to reflect changes in the scanner calibration.

STABLE SCANNER PERFORMANCE

It may be possible to reduce testing time if a scanner is deemed to be “stable”. That is, if it consistently passes “pre-test” and “post-test” comparisons when no adjustments to the settings have been made. Under these

circumstances, the “post-test” can be eliminated, as long as the last scanned document is checked for acceptability.

The minimum testing frequency should be a “pre-test” at the beginning of the work cycle and a “post-test” at the end of the work cycle.

DOCUMENT GROUPING

It may not be practical to do test runs before and after every document is scanned. A compromise can be reached if the source documents are grouped by similar types. If the same scanner settings would normally be used for all documents of a specific type, a test can be run before the batch and again after all the documents are scanned.

If the “pre-test” and “post-test” results match and the scanner settings have not been changed during the scanning process, the scanner can be assumed to have been working properly. If the test results do not match, the batch must be examined to determine which documents need to be reprocessed.

9.7 Test Logs

Logs should be maintained for all test runs. The log should include scanner settings. These allow management to confirm that the test runs are being performed and can help technicians identify potential problems. For example, if the scanner settings are gradually changing over time, these would be shown in the log, and may indicate that a light source is failing when reviewed by a technician.

The log should include the following information and should be recorded every time a test run is made, not just for the final test run. The additional information can be useful to technicians.

- Date and time of the test
- Who performed the test
- Scanner settings
- Type of documents scanned
- Test results

9.8 Creating Test Images

Test images can be created from the batch, or a “typical set” can be used. A typical set will not indicate whether or not the proper settings were selected for the document types, only if the selected calibration is right for that class of documents.

By using document types that are specific to the project, the operators will have both a test image that is similar to the images they normally work with, and information about what scanner settings are most appropriate for that document type.

For more information on how to determine a standard set, see Section 6: Document Survey.

9.9 Test Procedures

SCANNER SET UP

The scanner should be set to the specifications listed on the quality reference. However, an exception may be made in the case of a scanner that is slowly “drifting” in settings (the test log will indicate what type of drift is occurring). In this case, the last “good” settings in the test log can be used. This should only be a temporary solution until the scanner can be recalibrated.

Do not make any adjustments to the setting when running a post-test.

The test documents should be aligned the same way as non-test document. The test is defeated if, for example, the test documents are place by hand, but the non-test documents are loaded through a document feeder.

The same procedures used for scanning non-test documents should be used for scanning the test documents.

If a screen with sufficient image resolution is available, it may be used to determine the accuracy of testing, until the final settings are determined, at which time a hard copy can be printed. Screen examination should only be used if the screen resolution is good enough to clearly seen all important elements of the document and the image can fit completely on the screen.

Printing “pre-test” and “post-test” targets generally provides an easier method for comparison. These images should always be printed for the last set of test, and stored with the test log so that technicians can easily reference them.

PRINTER PROBLEMS

When there is a discrepancy between hardcopies it is essential to identify if the error is related to the scanning of the original image or the printing of the scanned result.

An electronic copy of a target quality reference image should be kept on disk. This image can be used to verify that a printer is working properly. The image can be printed and compared with earlier prints. If the prints match, the printer can be assumed to be working properly.

There are, of course, other areas of potential failure, such as the storage system. However, if the storage system were to fail, it would generally result in a partial, broken, or non-existent image. It is also likely the image could not be retrieved at all. Generally, if there are discrepancies in the final image, they will not be due to storage errors.

Once a problem has been identified, all printer and scanner settings should be recorded on the hardcopy, and saved for the technician to review.

TEST TARGET PHOTO COPIES

All photocopiers [22] introduce distortions of some type. Photocopying the target will destroy the usefulness of the size, placement, half and black tones areas of the test target.

Photocopies of test targets should, therefore, not be used as test targets.

9.10 Visual Inspection

INSPECTION OF DIGITAL IMAGE

A quality control program should be conducted throughout all phases of the digital conversion process. Inspection of final digital image files should be incorporated into your project workflow. Typically, master image files are inspected at the time they are created and/or via CD batch or online after the scanning and indexing are complete.

Depending on your project, you may want to inspect 100% of the master images or 10% of the files randomly, for example. It is recommended that images be examined as part of the scanning process, to minimize the risk of introducing images of substandard quality into the digital collection. A second inspection can be done, post-production, if resources permit.

The more clearly a project’s quality control procedures are documented, the easier it will be to define what constitutes an unacceptable defect in an image. Images should be inspected while viewing at a minimum pixel

ratio of 1:1 or at 100% magnification. The images may also be inspected at higher resolution levels, though when magnified in software, some interpolation may occur if the magnification is not a multiple of the original (for example, a 2x2 image viewed a 200% becomes a 4x4 image, with each original pixel taking up four new pixels, but when viewed at 150% each new pixel takes up 2.25 pixels, so clearly some manipulation by the software will be required in order to render an image that only uses full pixels).

Quality should be evaluated both subjectively by project staff (scanner operator, image editors, etc.) through visual inspection and objectively in the imaging software (such as using targets). The viewing environment for visual inspection of images is also important: monitors should be calibrated, and the room should be dark or at least free from bright lighting, sunlight, or glare.

Things to look for [13] during visual inspection may include:

- Image not the correct size
- Image not the correct resolution
- File name is incorrect
- File format is incorrect
- Image is in incorrect mode (i.e., color image has been scaled as grayscale)
- Loss of detail in highlight or shadows
- Excessive noise especially in dark areas or shadows
- Overall too light or too dark
- Uneven tonal values or flare
- Lack of sharpness or excessive sharpening
- Pixilated
- Presence of digital artifacts (such as very regular, straight lines across picture)
- Moiré patterns (wavy lines or swirls, usually found in areas where there are repeated patterns)
- Image not cropped
- Image not rotated or backwards
- Image skewed or not centered
- Incorrect color balance
- Image dull or no tonal variation

INSPECTION OF ORIGINAL HARDCOPY IMAGES

Some visual constraints can result in presentations of image garbage on your computer monitor. If the original image has fading, blurring, or “bleeding” of lines these will be transferred to the resultant digital image. The original should be compared side by side with the digital image.

Problems with the original [20] can also confound OCR processing. Characters that are touching or fuzzy may present problems. Strikethrough or other non-typical markups may also cause trouble. Unconstrained or hand printed characters should also be examined thoroughly in the OCR output.

Some OCR systems will be able to identify areas where text recognition is below some certainty level, making it easier to identify areas that may require manual intervention.

10 Sample Standard Tests

10.1 Standard Test Target: IEEE Std 167A Series

The IEEE Std 167A series was originally developed to test facsimile performance, but it can be used effectively for scanner testing. A sample is shown below (167A-1987). The notable features are the continuous tone gray-scale photo, varied alphanumeric characters and fonts, and resolution markers.



Figure 10-1 Standard Target: IEEE Std 176A-1987

There are several documents in the IEEE Std 176A series. Here we look specifically at the IEEE Std 176A-1987 specifications.

Field technicians [22] also commonly use this document.

GREY SCALE

The gray-scale test is useful for determining the point at which a system treats gray as black. As such, the original must be used, not a photocopy, since photocopiers will have their own thresholds that determine when gray is seen as black.

Photocopies of test targets should never be used. Only the unaltered original should be used in testing.

Grey scale patterns #7 and #8 are 15 step reflection density areas. Pattern 6 is a continuous density wedge. The values of the steps on the step wedge are given in the pattern description sheet that accompanies the IEEE Std 167A-1987 target. The primary use of these gray scales will be to show the threshold area at which a scanner decides that an area is black instead of white. Adaptive thresholding may result in different thresholds when going from black to white, instead of white to black.

Once the threshold settings are properly adjusted, the test target should be run again. Observe where transitions points occur. When comparing “pre-test” and “post-test” results, these breaks should occur in the same location.

RESOLUTION

NBS patterns are used to measure the point where black and white lines are not longer individually distinguishable, but instead appear as a blurred grey line. The value shown on the chart is called the “modulation”, and is calculated from normalized values of the black and white lines (where all black is zero and all white is 100). The modulation is calculated using the formula $(W-B)/(W+B)$ where W and B are the normalized values of the black and white areas. A modulation of one indicates that the lines are completely distinguishable, and zero indicates that they are completely indistinguishable. For comparison, a modulation of about 0.2 is the point where human eyes can no longer clearly distinguish between the black and white lines.

Use of the NBS 1010A resolution charts gives one a single point value, the highest resolution that is humanly discernable in the image.

The use of NBS 1010A bar type target resolution patterns are not recommended for scanners with a fixed aperture of less than 600 dpi.

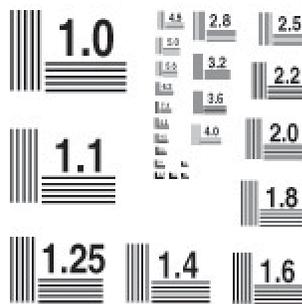


Figure 10-2 NBS Target

The Pestrecov Star pattern is composed of tapered lines over 360°. At the center the lines are very narrow, and will blur at varying distances from the center. The distance of the blurring from the center is an indication of the overall resolution of the system. The shape of the blurred area is a function of the direction of the scan. The size

of the blurred area may be a function of the threshold. The following figure [34] shows the Pestrecov Star with resolution blurring.

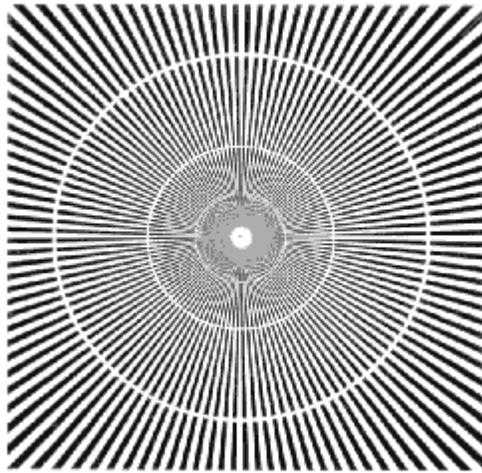


Figure 10-3 Resolution Blurring in the Pestrecov Star

Pattern #13 (the triangulated section with multiple trace rays) may also be used in the same way as the Pestrecov Star pattern, but in one dimension only. A moiré pattern will be produced at point approximately equal to the scanner resolution. The previous figure shows patterns emerging for resolutions of 50,100, and 200 LPI¹¹.

Several examples of moiré patterns [28] are shown below. These are all generated from the same original image, but taken at a different scan rate. The pattern occurs because different scanners will cross the threshold between white and black a different points. The size of resolution and the size of the line, as well as alignment and resolution will determine the final pattern. Moving the target on surface often will generate a similar, but different pattern.

¹¹ Lines per Inch: Frequency of occurrence of lines in a bar pattern. Also used for halftone/continuous tone printers.

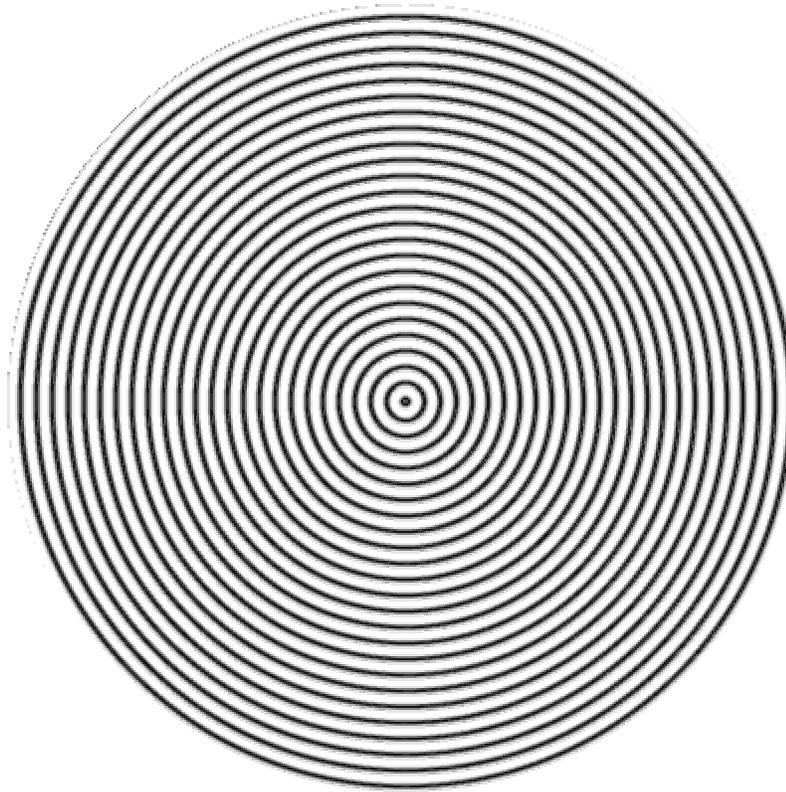


Figure 10-4 Concentric circle image with no Moiré Pattern

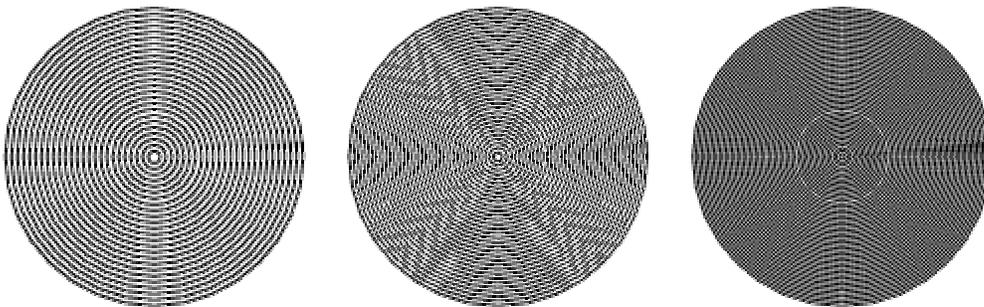


Figure 10-5 Same Concentric Circle Images with Different Moiré Patterns

LINEARITY AND RECTANGULARITY

Testing the linearity and rectangularity of a system is not normally a daily test. Rather, it is done after calibration to make sure the system is not distorting images.

To test linearity, carefully measure the inside lines of pattern #10. The length of line on the copy should match the length of line on the original. All lines should be straight and line on opposite ends should be of equal length.

If the rectangularity is correct, all opposite sides will have the same length and the length of the diagonals will be equal.

Note: This test may fail because of compression software or hardware [22], instead of the scanner.

TEXT

The test should allow the user to determine the smallest readable font size that can be distinguished by the scanner. The printers point values are shown next to each successively smaller alphabet.

ADDITIONAL TESTS

The test target also includes additional tests, which are described on the back of the target form [22]. The quality controller may also find these tests useful for some applications.

10.2 Standard Test Target: AIIM Scanner Target

The AIIM Scanner Target is an “ink on paper” [22] test target that simulates conditions that may cause scanner problems.

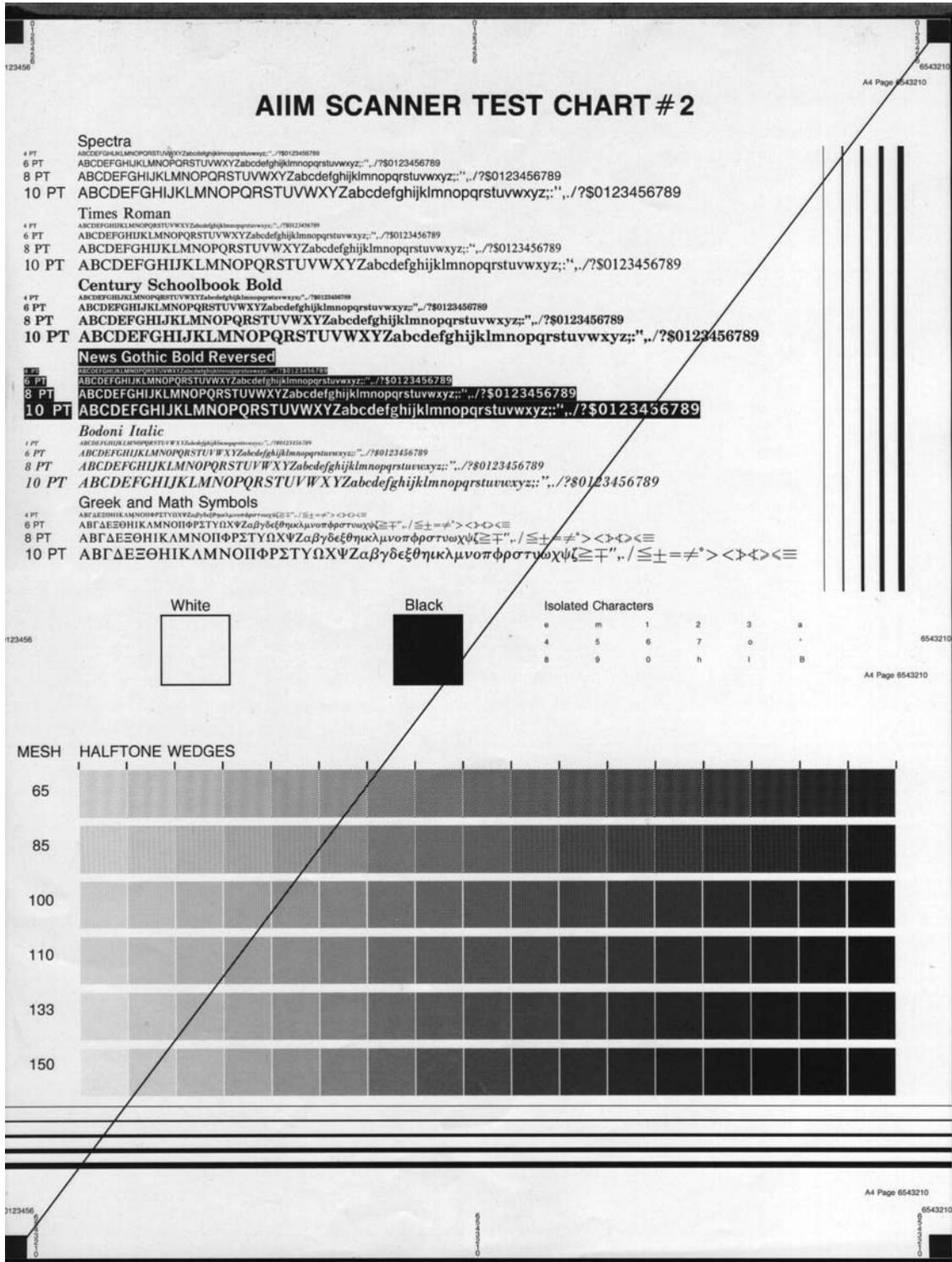


Figure 10-6 AIIM Target Sample

SCAN AREA

If the scanner has an adjustable scan area, check the area is of the proper size. At the corners of the target are black boxes that run off the edges of the target. If the boxes do not run off the edge of the scanned image, the scanner area may be too large.

Some scanners do not print the edge of the image. This must be considered when evaluating the printed target.

At the corners and in the center of each edge, a line of numbers runs of the edge of the target. If the digit zero is not visible in all corners of these images, the image has been clipped, and the scan area is either too small or the image was not properly aligned.

ALIGNMENT OF PAGE

If the scan area is of the proper size and the document is correctly aligned, the zero digits will be visible in all corners of the image and at the center of each edge. If the image size was reduced, look for the same numbers appearing in the same locations on all edges. If the numbers do not match, it indicates that the image was positioned off center or rotated.

Because this test is affected by target alignment, it can also be used to test the accuracy of automatic document loaders, after the initial tests have been run on the scanner.

TEXT

Text in a number of fonts and number of sizes is displayed in the text area. These represent some of the smaller sizes likely to be found in documents.

During initial examination of a scanner observe the small characters and punctuation to determine where potential scanner problems may occur. Look for legibility on the small characters as well as the presence of serifs. Examine fonts like the “News Gothic Bold Reverse”. These fonts may be susceptible to becoming filled with black by some scanners. It is important to know at what type size the scanner will lose the ability to distinguish between lower case letters like a, c, e, and o.

A properly adjusted scanner with a resolution of 300 dpi should be able to preserve this distinction on a 4-point type.

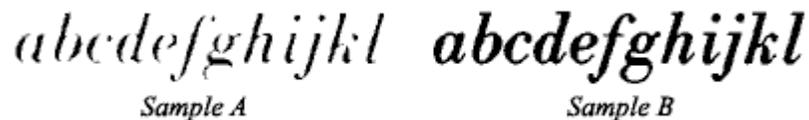


Figure 10-7 Two scans of the same image, on the same hardware with different calibrations

The stated size of type is in printer’s points. One point represents approximately 1/72nd of an inch. The point size of a given font is determined by measuring the distance from the highest ascender to the lowest descender. The actual height of any single letter is called the “x-height”. The “x-height” may vary by designer.

Legibility of a scanned letter will depend on the number of scan lines that fit within the “x-height” of any single letter, the quality of the original image, and the focus of the scanner. During a normal test run, examine the smallest letters that were recognizable on the quality reference made during calibration. If the scanner is properly adjusted, those characters will still be recognizable on the test run.

HORIZONTAL AND VERTICAL LINES

There should be five horizontal and five vertical lines on the page. Verify that the thinnest line is visible. This test confirms that there are no resolution “gaps” in scanner in either the horizontal or vertical. The gaps could occur if the resolution is too low, or there are other problems with the scanner.

Note that “stair-stepping” in the image is normal if the image is not parallel to the scan lines. This will be the case for all lines that are not positioned in direct alignment with the pixel grid, and will be especially true of line drawing that don’t follow a block format.

These artifacts are a byproduct of how scanners decide when a line falls within a pixel and when it falls outside. The best way to minimize this effect is to increase the resolution and color depth.

Stair-stepping (some times called “jaggies”) will be most noticeable in line drawings. If the drawings contain orthogonal lines along the horizontal and vertical axis, proper alignment can minimize the effect. However, most line drawings will have lines that don’t stick exclusively to the horizontal or vertical. The best scenario is to scan these images at high resolution with good color depth. The resultant tonal image will appear smoother.

OPTICAL CHARACTER RECOGNITION

Optical Character Recognition (OCR) is the process of using software to determine which characters appear in a digital image. The higher the tonal differences [38] between the character and background, the better the OCR software will be able to determine what the character is.

The highest tonal differences occur in pure black and white images. Although bi-tonal images make it easier for OCR algorithms to do character interpretation, they tend to make the images less “readable” to the human eye, as illustrated in the figure below.

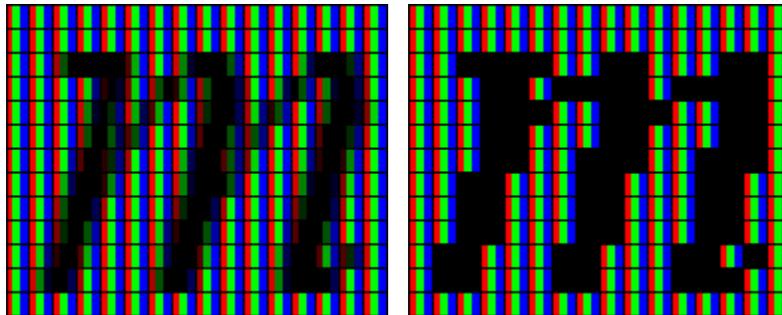


Figure 10-8 2-bit image (right) is a more accurate representation of the image, but the 1-bit image (right) is better for OCR

Therefore, it is recommended that a bi-tonal image be created, in addition to the full tonal master image, when OCR is required. The image may also be scanned at higher resolutions, if possible.

OCR can be timesaving when large amounts of text are involved, but it can also be prone to error. Older documents, especially those that are hand written, can pose a challenge to even the best OCR software, as can those whose writing is not aligned along any single axis.

DIAGONAL LINE

The diagonal line across the target is a test for uniform transport movement. The line should be smooth and straight within the capability of the scanner and recorder. Breaks in the line may indicate that the mechanical transport is not working smoothly or is being forced to pause and restart.

ISOLATED CHARACTERS

The isolated characters simulate a page number or mathematical equation. Because of the large white space around each character, some scanners will see the characters as large dirt specs [22], and eliminate them. Some scanners will fail on the degree symbol and display it as a solid dot.

BLACK AND WHITE AREAS

The black-and-white areas provide solid areas for density checking using a densitometer. Normally, visual examination is sufficient to determine if the white area is clear and the black area is solid.

Failure to show the black area as a solid black is normally a printer problem, and should be checked using the reference image.

Densitometric values for images that are output using toning processes vary significantly from the silver halide-based systems [22]. The user should be aware that measurement of toned images could be unique to that particular system. Presently, there is no standard for comparison of these measurements.

HALF-TONES

Half tones pose a problem for most scanners because there are only a small number of scan lines across each half tone dot. This can result in a moiré pattern that will depend on a variety of factors such as scanner resolution, half-tone mesh (dots per inch), and the angle and position of the target. Currently there is no way to eliminate moiré patterns, other than increasing the scan resolution (typically above 1000 lines per inch).

Given that most scanners will not reproduce half-tone dots perfectly, the evaluation criteria are the lightest (smallest) half-tone dot the scanner will recognize, and the darkest (largest) half-tone dot that the scanner will not see as a solid black dot.

On initial investigation these values should be noted, because if a half tone is scanned in a very light or dark area, the scanner may see only the solid area and lose the detail.

For the user this means that to capture all of the detail, the scanner typically has to be adjusted for each picture. If a very light area and a very dark area appear on the same document, the areas may have to be scanned separately.

The half-tone mesh (measured along a 45 degree angle) varies from publication to publication. A rough comparison is given below:

Newspapers	65-85
Technical Documents/Manuals	85 to 110
News magazines	110 to 133
Art magazines	150 and higher

During a normal test run, examine the lightest box visible and darkest box that is not solid black for each half-tone mesh. These should be the same on the quality reference output. The threshold setting of the scanner will have a direct impact on this test.

OTHER AIIM TESTS

Scanners can produce degraded imagery, producing images at much less than the default resolution, e.g., digitized images produced at 110 dots per inch (dpi) rather than at a default 240 dpi. One key industry standard for measuring scanner quality, ANSI/AIIM MS 44 with X440 Test Targets, measures 14 possible scanner mechanical and functional problems that can impact image quality, such as skewing, resolution, image contrast distortion, degradation, and image area exclusion.

10.3 Standard Test Target: RIT Process Ink Gamut Chart

The Rochester Institute of Technology (RIT) Ink Gamut Chart represents the range of colors that can be printed using standard process inks. This is the type of color printing which is found in most newspapers and magazines.

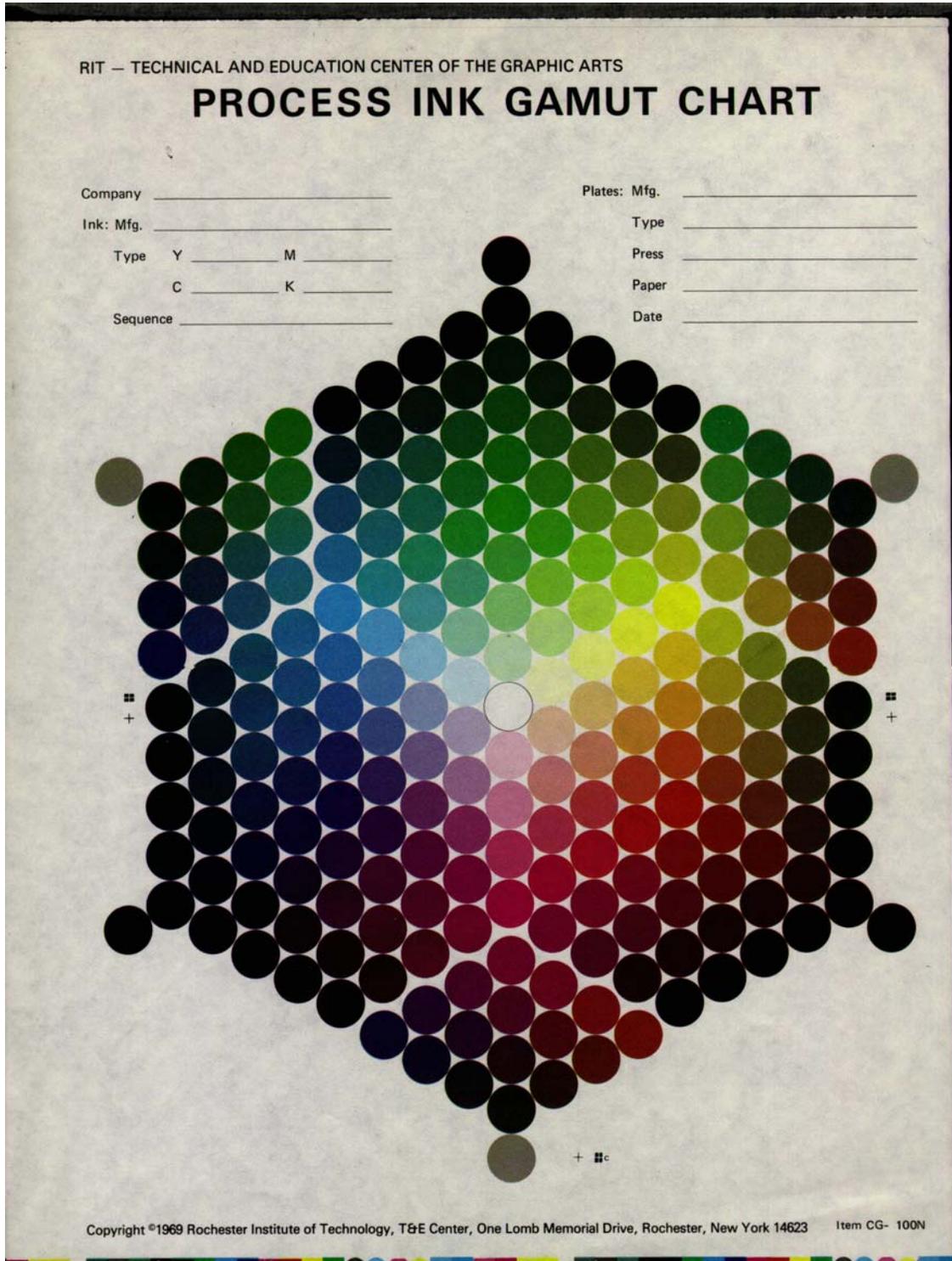


Figure 10-9 RIT Process Ink Gamut Chart

The process of ink gamut testing is an important one to run, even if color scanning is not normally done. The reason is that this target is sensitive to changes in the threshold setting of the scanner [22] and will show slight changes easily.

COLOR BLINDNESS OF SCANNERS

Most scanners are color blind in some color, typically the color of the light source they use. A scanner using a red laser light will not see some colors of red ink. A scanner using a blue CRT will typically not see some shades of blue ink.

Most scanners will resolve each scan site (pixel) into either black or white. This means that for every spot on every scan line, the scanner will make a decision, based on how much light is reflected from that site, whether to call the spot “black” or “white”. Black areas will be identified by the locations where the target absorbs the light from the source, and white areas will be identified by the areas that reflect the light.

Areas that reflect in the color of the scanner light source will reflect the same amount of light as white sources, and so will appear white.

Scanning and printing the gamut chart will result in a large number of black circles, some circles in various shades of gray with moiré patterns, and some circles not printed. The gray circles are due to the overlapping of the four different colors typically used to print color materials. In general, a color scanned on the gamut chart will be scanned in the same way on other scanned documents.

SCANNING COLOR PICTURES IN MONOCHROME

If color pictures are scanned on a monochrome scanner, the results will usually be poor. This can be seen from a quick examination of the gamut target scanned on a monochrome scanner. Most of the colors will scan as black, which will result in a black-on-black digital image, and a loss of visual data.

This high loss of visual information is unacceptable for digital master images. If a document collection contains color documents, monochrome scanners cannot be used. Some sources suggest using color scanners, even when documents are exclusively monochromatic, because the resultant grayscale tones can alleviate jaggedness that can occur in purely monochromatic scans.

10.4 Standard Test Target: IT8.7

This is another common test target, available in a variety of sizes, formats and reflectivities. The Kodak version is shown in the following figure.

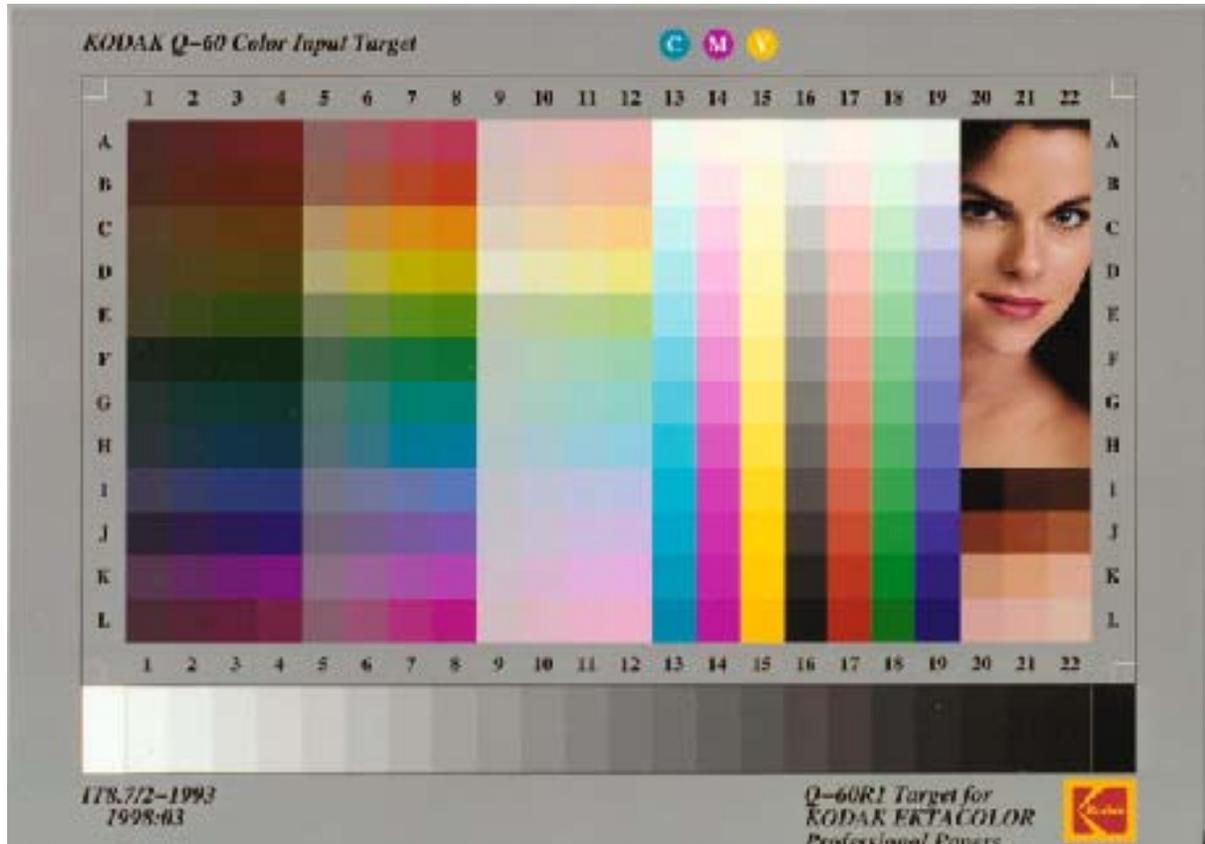


Figure 10-10 IT8.7 Test Target

COLOR ACCURACY

This target is intended for use in comparing color accuracy from the scanner with the true image. This target is often sold with software for automating the comparison. The colors in the target can be represented in 24-bit RGB color. The software then checks which RGB values is present in the scanned image and compares it against the *expected* value.

Alternately, any pattern of color swatches can be used, provide the RGB values are know explicitly for each swatch. These values can be checked using most image editing software (e.g. Adobe Photoshop).

Targets should be rotated and the test repeated, to determine if the scanner performs consistently for all areas of the scanning surface.

Remember that color targets are made with organic dyes and that these dyes breakdown as they age. Therefore over time the charts lose their accuracy.

11 Image Types

11.1 Image Types

The initial digital image does not necessarily have to be the only image created from the document. Additional images can be created for different tasks. In general, a single scan is used to produce a digital image of the highest quality associated with the project. This initial high quality image is called the “digital master” image.

Most digital archivers [13] recommend a primary master image be created which can be supplemented by an access image and/or a thumbnail image. The primary reason for creating supplementary images is to improve access to information in the digital master.

The needs of the project and the user base will determine which types of secondary images should be created.

For example, if records are to be accessed over the web, a low-resolution thumbnail image might be created from the original, because thumbnail images will load faster into a web-browser and allow the user to visually survey multiple images without having to load the entire master image file.

A second example might be enhanced images, in particular those created to clarify portions of the original image.

In a well-indexed system, the secondary images will all contain references to the original master image. This will allow users not only to see all derived images associated with a master image and move between the master and derived image, or between multiple derived images easily.

11.2 Digital Master Images

Digital master files are the images created from the initial scan of the physical document. The digital master should represent as accurately as possible the visual information in the original object. Digital images should be created through the direct scanning or imaging of the original physical document.

The primary function of digital master files is to serve as a long-term archival record and as a source for derivative files. A digital master file may serve as a surrogate for the original, may completely replace originals or may be used as security against possible loss of originals due to disaster, theft and/or deterioration.

Master documents represents as closely as possible the information contained in the original. They are generally uncompressed and unedited versions of the original image, which will serve as the long-term source for derivative files. Ideally, master files will be able to serve as surrogate for the original.

There are compelling preservation, access, and economic reasons for creating an archival-quality digital master image: it provides an information-rich, unedited, research quality surrogate, and ensures rescanning will not be necessary in the future. A high-quality master image will make the investment in the image capture process worthwhile.

In some cases the original object cannot be digitized directly due to its size or other attributes, it may be necessary to use a photographic intermediary or create a single large master from multiple, smaller masters. Anytime the master is not directly related to the complete original physical document, care should be taken that the exceptions are well documented and that the final master image represents the original object as accurately as possible.

Since the master will, ideally, be equivalent to the original document, it **should never be saved in a lossy image format**. Lossy compression techniques, such as JPEG, should not be applied to master files. Instead, lossless file formats (e.g. the TIFF file format) should be used.

Because a master image is of a higher quality than most other images and should be stored in a lossless file format, it will tend to have a relatively higher file size. The larger file size can make the image difficult to work with, both in terms of on-line storage and access over a network.

Many digital imaging projects [13] circumvent this problem by using the high-quality master as an archival image which is stored off-line. Multiple derive multiple versions of the image in smaller sizes or alternative formats are then created for on-line access. Derivative images are smaller files that can be more easily stored online, accessed over the network, and viewed through web browsers.

Digital master images should always be saved before any derivative images are created.

Since user expectations and technology change over time, a digital master must be available and rich enough to accommodate future needs and applications. The master image should be the highest affordable quality; it should not be edited or processed for any specific output; and it should be uncompressed. Intensive quality control should be applied in creating master image files.

The specifications for derivative files used for image presentation may change over time; digital masters can serve an archival purpose, and can be processed by different presentation methods to create necessary derivative files without the expense of digitizing the original object again. Because the process of image capture is so labor intensive, the goal should be to create a master that has a long useful life. The life of an image can, in theory, be indefinite. However, given historical changes in the field of image technology, it seems likely that the lifetime of an image, that is the duration for which it may be used without requiring some degree of maintenance or migration, is limited. There is also the possibility that as resolution and quality capabilities increase there may be a justifiable need to rescan images.

Bearing these considerations in mind, collection managers should anticipate a wide variety of future uses, and capture at a quality high enough to satisfy these uses. In general, decisions about image capture should err towards the highest quality.

FILM-BASED FORMATS

Normally, digital masters should record the visual appearance of the original image in a well-defined way, so that later users can knowledgeably interpret and manipulate the data, and so that capture technicians can follow consistent procedures for capture and quality control. For most kinds of documents and flat images, this can be accomplished by defining the desired digital values for standard targets included in the imaging. However, originals in other formats such as slides and negatives present different challenges.

Negatives are distinctly different from most images, because the visual appearance of the negative is normally of little interest; instead, users are generally interested in the positive image created from the negative. Both the methods and materials for making negatives have changed repeatedly over the years, as have the materials and practices used in making prints from them.

Planners may choose to express their capture plans in terms of the properties of the negative (e.g. transmission density), or try to describe a positive master format to be captured directly from the negative. Either course has pitfalls. While the first may seem like the safer approach, fully describing a negative will probably require more data than 8 bits per channel can carry, because negatives routinely record a much wider range of tones than media such as prints.

It would be advisable for the planners to fully consider each step of the image transformation, from the densities in the negative, to raw scanned data, to digital master, to derivative products, making sure that enough of the right information is available in the master to satisfy the needs of derivative production.

Also, as an aside, it's useful to remember that even black and white negatives often contain color information such as stained regions; scanning such a negative in color often reveals that one color channel emphasizes the effect of the stain, while another color channel may hide it.

Planners contemplating a project to scan color slides, some of which are faded, may find they need to make choices about whether to try to correct for the fading at the time of scanning, or to store a "faded", realistic digital master and then possibly produce a "restored" derivative to satisfy a need for an unfaded product. The second method would be favored because it provides for both a realistic and a "restored" version, and offers the possibility that different, better "restored" versions can be created in the future, regardless of any ongoing changes in the condition of the slide.

However, in some cases the scanner and its software may be better able to correct the faded color channels at the time of scanning by tailoring the information gathering to the levels of each dye remaining. The purposes and priorities of the project can help make the necessary choices: if a purpose of the digital project is to give lecturers a tool for selecting lecture slides for projection, or to record the condition of the slides, then the project will require a product that faithfully represents the faded condition of the slide. Experimental trials can reveal whether the "restoration" is better if performed at the time of scanning for a particular scanner and level of fading; in some cases multiple scannings and multiple masters may be the solution.

Good indexing and descriptive metadata will be the key to informing users about what type of "original" they are using. Creating both an unaltered "true" original and an enhanced original, then linking them descriptively, will let the users choose which image is most appropriate for their needs.

Negatives sometimes depict other documents or images (e.g. photos of photos, etc.) Again the question of whether the digital master should strive to represent the slide or the original the slide depicts come up. On one hand, if the slide itself is considered an original work, the digital master should probably represent the slide as accurately as possible. On the other hand, many digital image capture projects involve a film intermediate: the document to be captured is first photographed on film, and then the film is scanned to create the master file. In this case, the film may be seen as a vehicle for recording the tonality of the original document and the scanning of the film may be adjusted so as to correct for the color and tonal errors introduced in the filming.

The inclusion of a grayscale or color swatch, photographed together with the work of art, can make objective corrections possible.

FADED DOCUMENTS

Digital capture of faded documents presents many of the same challenges as scanning faded slides: is the objective to show the appearance of the document in its present condition, to make it maximally legible, or perhaps to depict it as we imagine it appeared as new?

Ordinarily the digital master would be made to depict the document as it exists, and the master would then be processed to create the legible or "reconstructed" derivatives. However, in some situations faded information may be better captured using extreme, non-realistic means such as narrow-band light filters, or invisible wavelengths such as infrared.

In these cases, multiple captures and multiple digital masters may be appropriate. Anytime multiple masters are made, or a single master is made from multiple smaller masters, the indexing and metadata should track these exceptions.

MICROFILM

Microfilm is a photographic medium designed not for natural, realistic tonal capture but for optimal legibility. A project to capture digital images from microfilm taken of manuscript originals might naturally choose to emphasize legibility over tonal accuracy in its masters and derivatives since the microfilm intermediate is already inclined that way; this would be an important consideration in choosing whether microfilm is an appropriate source for scanning for a particular purpose.

11.3 Access Image

It is generally recommended that three versions of an image be created: a master image, an access image, and a thumbnail image. A higher resolution access image may be created depending on the need to detect detail in the image.

The derivative images should be **created at the time the master image is created** during the scanning, but after the scanner calibration has been verified. See Section 5: Workflow for information on when derivative images should be created. Creating derivative at the same time the master is created ensures that the derivatives are available at the same time the master becomes available and makes creating the indexes easier, because all of the related digital images are created together, which eliminates the need to go back into the indexing system at a later time to link the master with the derivatives.

Note: Additional derivative images can be created at later times, as needed. However, any derivative images that are part of the project plan should be created as soon as possible after the master image is created.

Access images can be used in place of master image for general web access. Generally, they are scaled to fit within the viewing area of average monitor.

The smaller file size used for access images (compared to master images) provides faster download time through the network. This should be given extra consideration if a significant portion of the projected user base will be accessing the images through a slow connection (such as through dial-up). Access files are normally compressed to further improve access speed. They are usually stored in a compressed, lossy file format (e.g. JPEG).

Access images are the images that will actually be used by the end user, so they should be of a sufficient quality that then can be used for general-purpose research.

11.4 Thumbnail Image

Thumbnail images are small, low-resolution images intended to give users a quick preview of the full image. They are often presented with the data record (metadata) associated with the main image.

They are designed to display quickly online, there by allowing users to determine whether or not they want to view the associated access image. Their small size allows multiple images to be displayed within a single web page or other viewing interface. This is especially effective for displaying the results of a search or grouped images.

Thumbnail images are usually stored in a lossy, compressed file format, typically as GIF or JPEG file.

If the image consists primarily of writing (e.g. text, musical scores, etc.) thumbnails might not be suitable. At a reduced scale and resolution, there may not be much discernable difference between the thumbnail images of similar master images.

11.5 Enhanced Images

In order to ensure that a quality digital image is produced, it may be necessary to artificially enhance the original image. Image enhancement, if done correctly, can improve the usability and information content of a digital image.

However, because an enhancement represents a fundamental change in the original digital image, there is also a potential for the loss of data.

It is recommended that the original scanned image be preserved as the digital master (assuming that the digital master passes the other quality assurance tests) and that the enhanced image be archived along with the original. This provides the additional benefit of being able to use other enhancement algorithms in the future, which may produce better results, or enhance other features not present in the initial enhancement.

Some collections [12] will need to perform image processing on files for purposes such as removing blemishes on an image, restoring faded colors from film emulsion, or annotating an image. For these purposes we strongly recommend that a master be saved before any image processing is done, and that the enhanced image be used as a high-resolution derivative to generate further derivatives. In the future, as we learn more about the side effects of image processing, and as new functions for color restoration are developed, the original master would still be available.

SMOOTHING FILTERS

Smoothing filter image enhancement algorithms are designed to remove speckles and background texture (for example, highlighted fields) from images, which can improve readability. An example of how a smoothing filter can change an image is provided below. The figures [20] below show the original digital image (left) and the image after have been artificially enhanced by a smoothing filter (right).

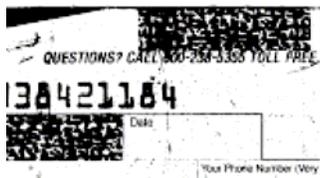


Figure 11-1 Before Smoothing

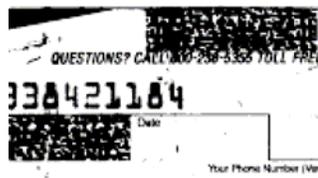


Figure 11-2 After Smoothing

EDGE DETECTION

Edge detection image enhancement algorithms can improve the shapes of scanned characters and graphics, and restore faded and poor-quality originals. This algorithm can enhance very faint line strokes and generate well-shaped solid characters without introducing speckled noise, a common side effect of other algorithms that use local contrast enhancement. The sensitivity of an edge detection algorithm can either be set to a fixed value by the user or dynamically computed by the board.

An example [20] of how edge detection can change an image is provided below. The figure on the left shows the original digital image and the figure on the right shows the image after it has been artificially enhanced by an edge detection algorithm.

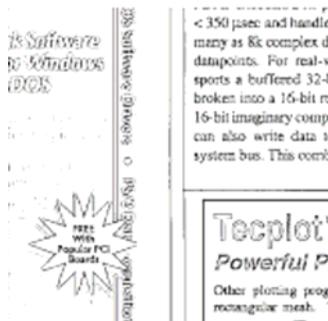


Figure 11-3 Before Edge Detection

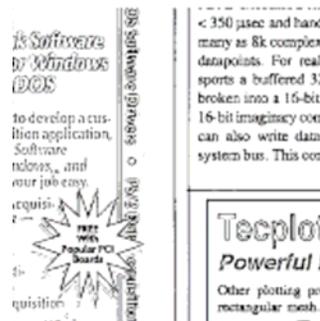


Figure 11-4 After Edge Detection

Edge detection and many other image enhancement filters are available in common commercial image editing packages (e.g. Adobe Photoshop) and with some scanner software.

DESKEWING

If the vertical and horizontal axis of the original document is not aligned with the axis of the scanner, the digital image produced will be slightly “skewed”. The document axis will not appear vertical/horizontal when the image is viewed.

To correct this problem, sophisticated “deskewing” algorithms can be used to rotate the image back to a true alignment. Unless the image is rotated by a multiple of 90°, the location of a pixel will not be rotated to another exact pixel location; instead multiple pixels will usually overlap destination pixel locations. The algorithms are used to blend the pixels so that the image is preserved.

This blending can cause a loss of image information, so deskewing should not be used to alter the master image; rather, it should be used to create a new derivative image from the master image.

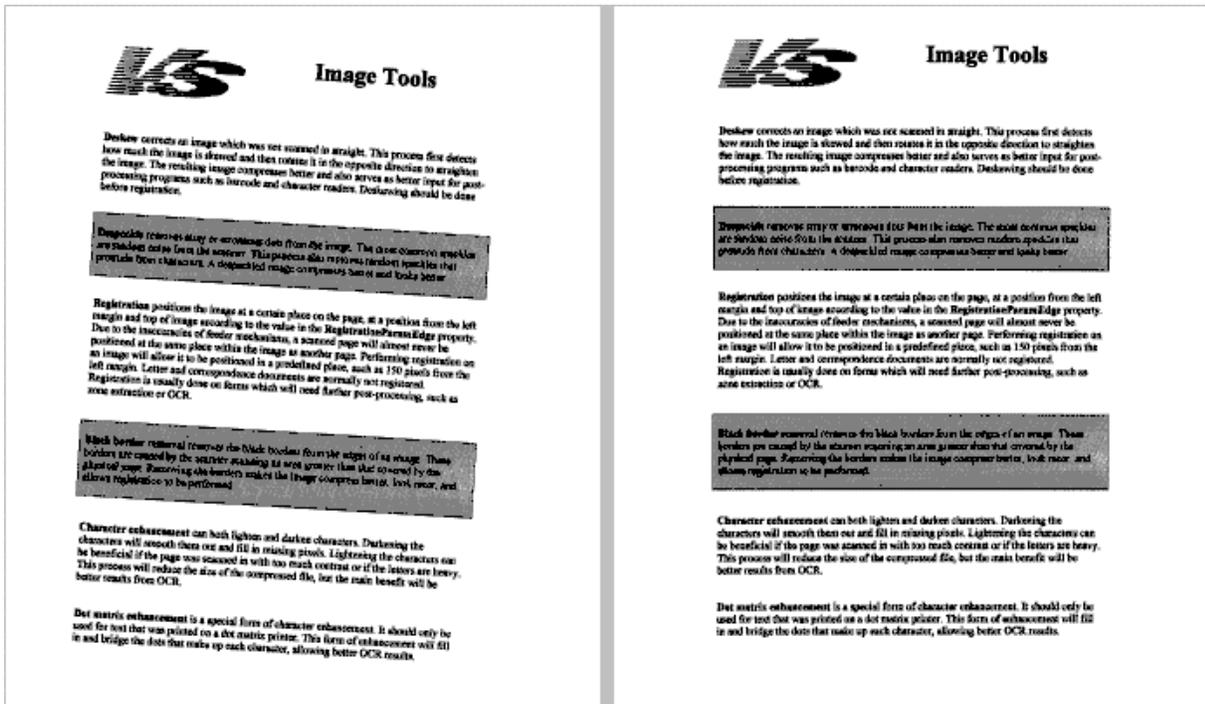


Figure 11-5 Example of Image Deskewing

Because the pixel content is being rotated, the image can suffer additional data loss (because pixels are rarely rotated so that they will be perfectly

11.6 Grayscale and Halftones

There is sometimes confusion between the concepts of grayscale and halftones. Halftone images use discrete dots of pure black in varying sizes to give the impression of a continuous grey-tone. Half-tone is traditionally related to printed images.

However, halftones are also used in digital images. The main difference between their analogue counter parts in the printed world is that digital images can not vary the size of the “dots”, because they are represented using individual pixels.

The following figure [34] demonstrates visually the difference between grayscale images and those that use halftoning.

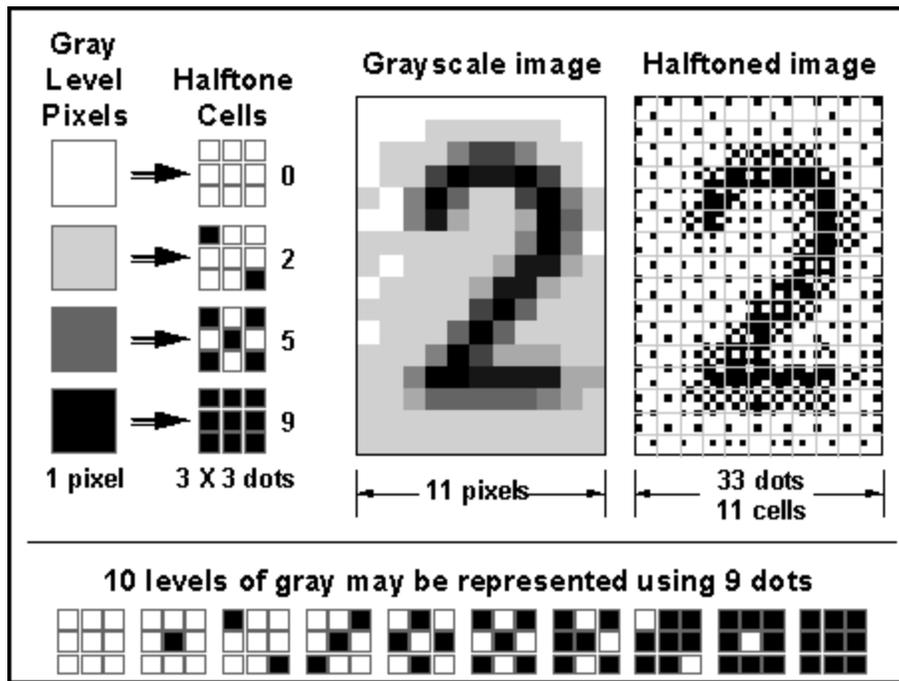


Figure 11-6 Graphical Representation of the Difference between Grayscale and Halftone Images

12 Metadata

12.1 Introduction to Metadata

Metadata is, literally, data about data. Metadata describes both the content of the data and how that data is organized, including descriptive information about the context, classification, quality and condition, or characteristics of the data. One of the key features of metadata is that it is self-describing, which makes it a good candidate for any type of data interoperability project.

No small or large scale scanning project can be successful without investing in the design and planning of metadata requirements and budgeting for the metadata capture. Metadata capture is best when accomplished “in-process” rather than planning for post-scanning collection.

For scanned documents, the metadata represents flexible information about both the original and electronic document.

Metadata is self-describing, thus making it possible to establish multiple data relationships and add new “meta-columns” to the data for specific maps. The collective metadata for a set of electronic documents can then be queried and summarized, just like traditional database fields.

Although this document focuses on digital metadata, not all metadata is in a digital format. Metadata already exists in a non-digital environment, such as legends on the physical map or paper catalogue information.

Metadata does not need to come from a single source, nor does it need to be created entirely at one point in time. Because metadata is self-describing, new information can be added to the metadata over time and from a variety of different sources.

A good object will have and be associated with metadata. All good objects will have descriptive and administrative metadata. Some will have metadata that supplies information about their external relationships to other objects (e.g. the structural metadata that determines how page images from a digitally reformatted book relate to one another in some sequence).

Some objects will have metadata embedded within them such as an encoded text with an XML¹² header; an image with a TIFF header. XML headers can use the tags to describe the contents of the XML body. TIFF headers can contain additional information, not related to image format, such as the document name, image description, scanner software, and author.

With others, metadata will be stored and managed separately.

Without special software, computers generally cannot use the informational content of a raw image file to search for or retrieve a specific image. Search and retrieval normally depends on some external form of indexing, which assigns specific metadata to each document. The metadata then contains the information that will be used in searching (e.g. author, recipient, date, title, and content keywords, etc.). The indexing, or metadata, can be simple or sophisticated, and is typically an electronic database that is linked to the images.

The indexing should also contain information indicating where the actual image is located as well as any related (e.g. master, derived, or enhanced images) documents.

Useful indexing requires careful planning and forethought in establishing what types of metadata will be collected. These requirements should be clearly established and communicated to project staff members before any actual imaging begins

¹² Extensible Markup Language

12.2 Advantages of Using Metadata with Archival Image Data

Metadata provides information about the structural relationships between different parts of a digital collection, including where and how the images were created from the original and where they are archived.

The more highly structured the information associated with digital records and images, the more that structure can be exploited for searching, manipulation, and interrelating with other records in the digital collection. Capturing, documenting, and enforcing that structure, however, can be a time consuming task that requires specific types of metadata, an explicit vocabulary, and clear set of project goals.

In an environment where a user can gain unmediated access to records in the digital collection over a network, metadata can be used to certify the authenticity and degree of completeness of the content within the digital image by establishing and documenting the context of the content of the image. The metadata should be able to identify (and exploit) the structural relationships that exists between itself and other digital images (by way of their metadata) in the digital collection. It can expose a range of intellectual access points for an increasingly diverse range of users and provide some of the information that an information professional might have provided in a physical reference or research setting.

Metadata can also be added to the image record that relates to the administration, access, preservation, and use of the records or the whole collection.

12.3 Typical Metadata Categories

ADMINISTRATIVE

Administrative metadata is used in managing and administering information resources. Typical examples might include: acquisition information, rights and reproduction tracking, documentation of legal access requirements, location information, selection criteria for digitization, version control (along with any differentiation between similar information objects), and audit trails created by record keeping systems.

DESCRIPTIVE

Descriptive metadata can be used to describe or identify information resources. Typical examples include: cataloging records, finding aids, specialized indexes, hyperlinked relationships between resources, annotations by users, and metadata for record keeping systems generated by records creators.

PRESERVATION

Preservation metadata is related to the preservation management of information resources. It might include how the image is stored, when the image should be migrated, and what the expected retention period should be.

Although the metadata is focused on the digital record, it may be appropriate to enter information about how the original physical object was preserved, as that information may have an impact on how the original physical document looked when it was first digitized.

TECHNICAL

Technical metadata is related to how a system functions or metadata behaves. Typical examples include: hardware and software documentation, digitization information, e.g., formats, compression ratios, scaling routines, tracking of system response times, and authentication and security data (e.g., encryption keys, passwords).

USAGE

Usage metadata is related to the level and type of use of information resources. Typical examples: exhibit records, usage and user tracking, and content re-use and multi-versioning information.

This information may also be automatically updated by the access software to provide imbedded access and usage statistics.

12.4 Typical Metadata Characteristics

DATA SOURCE

There are two primary types of data sources: internal and external.

Internal metadata is generated by the creating agent for an information object at the time when it is first created or digitized. Typical examples of internal metadata include: file names and header information, directory structures, and the file format and compression scheme.

External metadata is data relating to an information object that is created later, often by someone other than the original creator. Typical examples include: registration and cataloging records, and rights and other legal information.

METHOD OF CREATION

There are two basic divisions of how metadata can be created: automatic and manual.

Simply put, automatic metadata is generated by a computer. Typical automatic metadata creation includes: keyword indexes, and user transaction logs. In some cases OCR can be used to automatically add metadata to the record.

Manual metadata created by human operators of the scanning system. This encompasses the majority of metadata not related to the image creation and all of the expert metadata.

It is not uncommon for all metadata to be entered manually, although this can be a time consuming task. Repetitive tasks that may be error prone (creating catalogue numbers, logging the creation time or file name) can benefit greatly from automation.

METADATA NATURE

The nature of metadata is divided based on the expertise of the data creator. At one end of the spectrum is “lay” data, created by users who are not necessarily subject master specialists. The information contains “high-level” information (e.g. file system information or data intended to be presented through a web page). The “lay” data is usually entered by the person who created the digital record.

At the other end of the spectrum is “expert” metadata which normally includes specialized information about the object, detailed subject headings, or archival indexing. Creating “Expert” metadata can be more resource intensive than “lay” data (depending on the volume of data to be created); as such it is often added after the original digital image is created.

STATUS

Status describes the “permanence” of the metadata.

Static metadata never changes once it has been created. Static metadata includes: title, provenance, and date of creation of an information resource.

Dynamic metadata may change with use or manipulation of an information object. Typically this includes information like directory structure, user transaction logs, and image resolution.

Long-term metadata is necessary to ensure that the information object continues to be accessible and usable over the course of time. Long-term metadata may include technical format and processing information, rights information, and preservation management documentation.

Short-term metadata is mainly of a transactional nature.

STRUCTURE

Structured metadata is metadata that conform to a predictable standardized or non-standardized structure such as MARC¹³, TEI¹⁴ and EAD¹⁵, or local database formats. Unstructured metadata is metadata that does not conform to a predictable structure, such as note fields and annotations.

CONTROL SEMANTICS

Controlled metadata uses standardized vocabulary or authority form to provide values for metadata field. There are a variety of “libraries” of standardized terms and authorities that can be used. Some common authority sources are AAT¹⁶, ULAN¹⁷, and AACR2¹⁸[15]. These standardized libraries provide common terms that can be used for anything from authors’ names, to standard abbreviations or technical terms. Standardized vocabularies are usually collaborative efforts, whose definitions grow over time. Standardized vocabularies can make it easier for indexers to determine what values should be supplied when the metadata is created so that users will have an easier time locating and interpreting the information.

Uncontrolled metadata is metadata that does not conform to any standardized vocabulary or authority form. Typical examples of uncontrolled metadata semantics include free-text notes and HTML metatags, or any type of open metadata tags or values.

LEVEL

Collection metadata is metadata relating to collections of information objects, such as collection-level records (e.g., metadata formats specific to cataloging like MARC (not to be confused with MARKS, an older Army record management system) records or other finding aids), and specialized indexes. Item metadata relates to individual information objects, which often contained within collections, such as transcribed image captions and dates, and format information.

¹³ Machine Readable Cataloging. See [32] for more information on this standard. This should not be confused with the USACE MARKS database.

¹⁴ Text Encoding Initiative. TEI is a method of “marking up” text documents (using SGML DTDs), primarily focused on academic sources. See [33] for more information.

¹⁵ Encoded Archival Description. Also uses SGML DTDs for archival indexing and searching.

¹⁶ Art and Architecture Thesaurus which lists about 125,000 standard terms architectural, material, and archival information. See <http://www.oclc.org/news/announcements/announcement21.htm> for more information .

¹⁷ Union List of Artists Names. Supplies a standardized list of about 2500,000 artists and collaborator names. This is a compiled list to which the USACE could contribute. For more information see:

http://www.getty.edu/research/conducting_research/vocabularies/ulan/about.html.

¹⁸ Anglo-American Cataloging Rules, Second Edition. for more information see :
<http://www.library.cornell.edu/tsmanual/CIRM/AACR2.html>

12.5 Metadata Structure

INFORMATION OBJECT

One of the key components of the metadata structure is the concept of an “information object”. An information object is a digital item or group of items, regardless of type or format that can be addressed or manipulated as a single object by a computer. This concept can be used to refer both to actual content (such as digitized images) and to content surrogates (such as catalog records or finding aids). In this context, an information object is anything that can be addressed and manipulated by a human or a system as a discrete entity. For digitization projects, it represents the combination of the image file, indexing and metadata.

The metadata represents what is “known” about the image file, and the indexing indicates how the image file can be found.

The object may be comprised of a single item of information, or it may be an aggregate of many items. In general all information objects have three [15] basic features:

- Content
- Context
- Structure

INFORMATION OBJECT CONTENT

Metadata associated with a digital information object’s content specifies what type of information is contained within the object. This is essentially “what the object is about” and any intrinsic information contained within the digital object.

The content provides high level information about the digital object and is one of the primary areas of metadata that should be populated for successful searching.

The information described in the content metadata section can be highly varied. This section will, therefore, benefit the most from the use of a standardized vocabulary. A standardized vocabulary is a specific set of terms that can be used to describe technical, visual, and other components of a digital object.

The use of a standard vocabulary greatly improves the reliability of searching records because both metadata authors and users will be using the same terminology.

INFORMATION OBJECT CONTEXT

The context of a digital object is related to the objects creation and can also include information about the creation of the original physical document used to create the digital object. It includes the “who, what, why, where, and how” of where the digital object came from.

Context generally employs a more restricted vocabulary than content.

INFORMATION OBJECT STRUCTURE

“Structure relates to the formal set of associations within or among individual information objects and can be intrinsic or extrinsic.” [15]

The structure can be related to the specific metadata structural format, or the format of the digital image.

12.6 Metadata and Digital Archives

The information contained in digital archive metadata is used to add extra value to the digital image and provide archival information as well. This additional information describes the arrangement, tracking and otherwise enhances access to the digital image. The goal of metadata within a digital archive is to enhance the intellectual access to the information in the digital image.

Digital archive metadata includes indexes, abstracts, and catalog records created according to cataloging rules and structural and content standards such as MARC (Machine-Readable Cataloging format), as well as authority forms such as LCSH (Library of Congress Subject Headings) or the AAT (Art & Architecture Thesaurus). Collaborative bibliographic metadata initiatives have been in place for over 40 years. These collaborative lists can be made available to digital archive projects and users through automated systems such as bibliographic utilities, online public access catalogs (OPACs), and commercial online databases.

Keeping the context clear is what assists with identifying and preserving the information value of records and images over time; it is what facilitates the authentication of those records and images, and it is what assists researchers with their analysis and interpretation of the information that they contain.

Archival and manuscript metadata generally includes acquisition records, finding aids, and catalog records. Archival descriptive standards that have been developed in the past two decades [15] include the MARC Archival and Manuscript Control (AMC) format published by the Library of Congress in 1984 (now integrated into the MARC format for bibliographic description); the General International Standard Archival Description (ISAD-G) published by the International Council on Archives in 1994; and Encoded Archival Description (EAD), adopted as a standard by the Society of American Archivists (SAA) in 1999. While archival metadata has primarily existed in print form until recently, it is increasingly distributed on line through resources such as the RLIN AMC (the Research Libraries Information Network Archival and Mixed Collections file, Archives USA, and EAD-based archival information systems.

12.7 Metadata Specification Options

Among professional and cultural archivists (museums, libraries, etc.) there is a wide variation in which metadata standard is implemented. Many highly detailed metadata standards are now emerging (such as FGDC¹⁹, SDSFIE²⁰, EAD and the Australian Record keeping Metadata Schema (RKMS)) that try to address the mission-specific need of these diverse groups while still providing a mapping between common data elements.

DUBLIN CORE

By contrast, the Dublin Core Metadata Element Set (DC) identifies a minimal set of common metadata elements that creators or catalogers can assign to information resources, regardless of the form of those resources, which can then be used for network resource discovery.

Because the Dublin Core can be represented using either HTML or XML, it integrates well into World Wide Web solutions.

The Dublin Core specification uses one-to-one mapping of object to information, which means that an aerial photograph of location would be described as itself (i.e. the photograph) and not as the location. It also uses a minimization principle (sometimes called the “dumb down principle”) whereby qualified descriptives in the metadata can be used with only the value and not the qualifier.

¹⁹ Federal Geographic Data Committee. An outline of the standards defined by this group can be found at [30].

²⁰ Spatial Data Standards for Facilities, Infrastructure, and Environment. This CADD/GIS group provides standards for grouping geographic references and associating them with attributes. More information can be found at [31].

METS

The METS²¹ schema is a standard for encoding descriptive, administrative, and structural metadata which is designed for digital library collections. METS uses a standard XML schema language, based on the W3C²² standard. The XML standard eases integrations with web technologies, since XML is supported by most web browsers.

The METS was developed as an initiative of the Digital Library Federation. The standard is maintained in the Library of Congress, in their Network Development and MARC Standards Office.

XML

XML (the Extensible Markup Language) is based on SGML²³, but has been greatly simplified. Because it is based on SGML, it interoperates well with SGML and HTML, making it a popular method for organizing data that will be shared over the web. However, XML is not a web-specific technology, and can be used to mark up any type of document.

ARIMS

Army Records Information Management System (ARIMS), is not strictly a metadata system, instead it is a record management system that supports metadata using its own standardized vocabulary. This vocabulary could be extended to support USACE needs if the existing one is not sufficient.

One of the advantages of ARIMS is that it provides a structure indexing, metadata support, web access and other tools and is available without charge to the USACE through fiscal year 2009[42]. This system places emphasis on important records, provides secure management of electronic records and simplifies the process of identifying and preserving important records in any medium. Additionally, it provides management for the original hardcopy documents.

12.8 Metadata Workflow

Metadata creation and management can become a very complex mix of manual and automatic processes. Often different staff members will add different layers of information to the metadata at different times. For example, the staff member doing the scanning may be responsible for assigning a digital catalogue number when the digital image is first created, whereas a subject matter expert may be responsible for adding information about the content of an image after the scanning and quality assurance phases are complete.

The following figure [15] illustrates the different phases through which metadata information “objects” typically move during their life in a digital environment. As they move through each phase, the digital images will be assigned additional metadata that can be associated with the images in several ways.

This metadata can be contained within the digital image itself. The metadata could, for example, be imbedded in the header information of an image file (e.g. Dublin Core) or through some form of bundling (e.g. UPF²⁴).

Metadata can also be attached to the digital image through external bi-directional pointers or hyperlinks, such as by adding the metadata directly to the indexing database, or by including references to the metadata there.

In any instance where it is critical that metadata and content coexist, it is recommended [15] that the metadata become an integral part of the information object and not be stored elsewhere.

²¹ Metadata Encoding Transmission Standard. See <http://www.loc.gov/standards/mets/> for more information.

²² World Wide Web Consortium. A standards group for web technologies. <http://www.w3.org/>

²³ Standard Generalized Markup Language. A robust, powerful, and complex markup language that can be used to describe the structure and content of a document. It is an international standard (ISO 8879:1985).

²⁴ Universal Preservation Format. <http://info.wgbh.org/upf/>

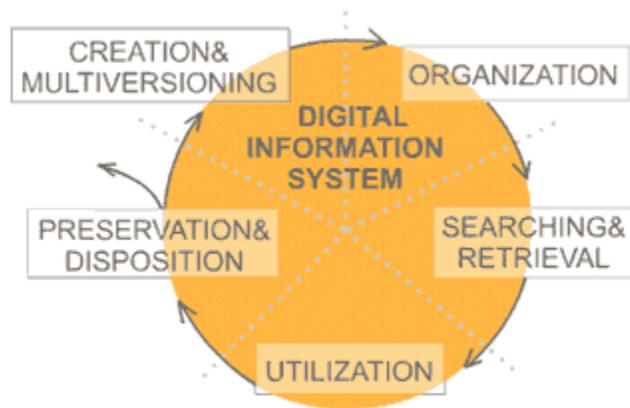


Figure 12-1: The Life Cycle of Objects Contained in a Digital Information System

CREATION AND MULTI-VERSIONING

Objects enter a digital information system by being created digitally or by being converted into digital format. In the case of a digital imaging project, this transition occurs during the initial scanning of the original physical image. Multiple versions of the same object may be created for preservation, research, dissemination, or even product development purposes.

Metadata often is added by the scanner that includes basic archival information, such as digital catalogue numbers, date of creation, where the image is archived (e.g. CD-ROM, network drive), etc.

ORGANIZATION

Digital projects should have a clearly stated metadata structural plan. They may chose to implement their own custom metadata system, or may use one of the many existing structural standards, but one the structure is chosen, all metadata should adhere to that particular standard.

Once the digital image metadata has been created, it should be moved into the organization of the overall metadata system. This may be done through either an automatic or manual process. Additional metadata for those objects may be created through registration, cataloging, and indexing processes.

SEARCHING AND RETRIEVAL

The main purpose of metadata is to provide users with flexible information about the content of images in the digital collection. For most users, searching the metadata will be the best way for them to uncover digital images that meet their requirements.

Metadata should be organized in a way that not only makes it searchable through the main organizing system, but ideally also tracks what types of searches are performed. Tracking how users search metadata will help project planners decide if changes need to be made to the metadata exposed to the search utility and/or the search utility itself.

UTILIZATION

Utilization metadata covers two areas related to the history of the record itself: how it has changed and how it has been accessed. Utilization primarily focuses on how the object has changed (e.g. when the object was modified or reproduced, when new versions were created), along with the “who, when, where, and why” associated with the changes. It also encompasses user annotations and rights tracking, where applicable.

Usage statistics may also be saved to complete the digital “paper” trail.

PRESERVATION AND DISPOSITION

At some point most digital records will undergo refreshing, migration, integrity checking, to ensure their continued availability. Digital records that are no longer necessary may be discarded. Metadata should document both preservation and disposition activities, as well as schedules for migration or disposal.

12.9 Unique Identification

A good digital image will be named with a persistent, unique identifier that conforms to a well-documented scheme. It will not be named with reference to its absolute filename or address (e.g. as with URLs²⁵ and other Internet addresses) as filenames and addresses have a tendency to change. Instead, the filename's location should be resolvable by using a reference to its identifier.

As an example, consider a bookstore. The location of a book on the shelves may change over time, but the ISBN number that uniquely identifies that book will always be the same.

How an image is identified determines how (even whether) it may be found and thus made accessible over both the short and longer terms. There are at least two approaches to the assignment of persistent and unique digital record identifiers. The first involves assigning identifiers that conform to a standard, and using applications that ensure that those names resolve to the object's filename and location.

A second, more local approach may be considered, when the application of national or international standards is beyond an institution's technical capabilities. This approach involves developing and maintaining a local scheme that uniquely identifies information objects, and mechanisms for ensuring that names resolve to file locations. Where local schemes are used they should be documented and documentation should be accessible.

A third way that is appropriate for Internet accessible images (and metadata) can be implemented by assigning PURLs (Persistent URLs) instead of URLs. The PURLs embedded in references to the object can then be resolved to the true image location by a server that contains tables mapping PURLs to URLs. Although the mapping tables must be updated when an object is moved, this degree of indirection facilitates maintenance by ensuring each PURL need only be updated once in a central location, no matter how many times it occurs in references.

12.10 Authentication

A good digital image can be authenticated in at least two senses. First, a user should be able to determine the image's origins, structure, and developmental history (version, enhancements, etc.). Second, a user should be able to determine that the object is what it claims to be.

For the purposes of research, users must be able to verify the authenticity of any records being used. There are some cases where verification takes on additional significance, as for example, with the networked representation of information that supplies evidence about important past or current events.

²⁵ Uniform Resource Locator. The name for a resource (web page, image, PDF document, etc.) on the web. For more information, check the following URL: <http://www.w3.org/Addressing/>

Typically, the information necessary for a user to determine an images origin, structure, and developmental history should be included with the metadata that is supplied for and about that image.

Techniques for determining the veracity of digital images may include digital signatures and water marking. Checksums and other technical routines that produce message digests are appropriate for images in virtually all formats. They help determine by analyzing the object's structure and composition whether it has been changed in any way since some particular point.

Checksums can be used to verify that an image has transferred successfully from one location to another, such as may be the case when moving the images from a hard drive to a CD-ROM after scanning.

12.11 Self-Describing

Collections and digital images should be described so that a user can discover important characteristics of the collection, including scope, format, and restrictions on access, ownership, and any information significant for determining the collection's authenticity, integrity and interpretation.

The collection description is one form of metadata that two purposes: it helps people discover the existence of a collection (whether they are end-users seeking materials relevant to their information needs, or other collection-builders looking for similar or complementary materials), and it helps users of the collection understand what they are looking at.

Self-describing data can therefore be used to establish hierarchical relationships with other metadata. As such, digital imaging projects can (and should) support collection metadata, with individual groups, topic areas, etc. providing a bridge to individual metadata documents.

To serve the first purpose, when possible, collections should be described in collection-level cataloging records available at the user level. Websites and individual digital objects can be cataloged any of a number of directories where collections can be registered (e.g. OCLC's CORC²⁶).

12.12 Intellectual Property

If all of the documents in a collection will be internal to the project and not shared with external sources, there may be no need to account for intellectual property rights. However, in general, a good collection should respect intellectual property rights. Where applicable, collection managers should maintain a consistent record of right's holders and permissions granted for all applicable materials.

Intellectual property law should be considered from several points of view in relation to any collection: what rights the owners of the original source materials retain in their materials; what rights or permissions the collection developers have to digitize content and make it available; what rights collection owners have in their digital content; and what rights or permissions the users of the digital collection have to make subsequent use of the materials. Viewed from any side, rights issues are rarely clear-cut, and the rights policy related to any collection is more often a matter of risk management than one of absolute right and wrong.

Although this will not necessarily affect all projects, it should be give some consideration to ensure that it can be eliminated from the project requirements without concern.

12.13 Usage Statistics

A good collection provides some measurement of use. Counts should be aggregated by period and maintained over time so that comparison can be made.

²⁶ <http://corc.oclc.org> is one of many cooperative online cataloguing initiative

Measures can include use counts ("x files retrieved"), user analysis ("this site was visited by x users from y different domains"), or "linked-to" counts ("this site is linked to by n other sites"). Since measures should be maintained over time, they take some resources to support, and the measures chosen should be designed to serve some purpose of the sponsoring project or organization. One common use is to attempt to justify resources devoted to a collection by volume of use, either generally or within a certain user population. Another use is to enlighten collection development policy. Metrics are also a tool in the evaluation of projects and collections. [17]

13 Collecting Metadata

13.1 Advantages of Using Metadata

INCREASED ACCESSIBILITY

Properly implemented, rich, consistent metadata can significantly enhance search efficiency. Metadata can also make it possible to search across multiple collections or to create virtual collections from information that is distributed across several collections, but only if the metadata structures are the same or can be mapped across each site.

Digital information systems and emerging metadata standards developed by different professional communities but incorporating some common data elements, such as Encoded Archival Description (EAD), the Text Encoding Initiative (TEI), and the Dublin Core are making it easier for users to negotiate between descriptive of digital images and the images themselves, and to search at both the record (i.e. image) and collection level within and across information systems.

Museum, archival, and library repositories do not simply hold objects. They maintain collections of objects that have complex interrelationships among each other and associations with people, places, movements, and events. In the digital world it is not difficult for a single object from a collection to be digitized and then to become separated from both its own cataloging information and its relationship to the other objects in the same collection. Metadata plays a critical role in documenting and maintaining those relationships, as well as in indicating the authenticity, structural and procedural integrity, and degree of completeness of information objects.

The combination of metadata and image indexing provides users with a powerful method of finding relevant digital images. The metadata will need to integrate with any custom cataloging or indexing.

EXPANDING USE

Digital information systems for museum and archival collections make it easier to share digital versions of their collection to users through networked access. Network access can make records easily accessible to users who might otherwise have to travel to the location of the collection. However, networked access also presents a new challenge in that the records have to be made accessible in a way that is available to the use – that is, through the access software instead of a collection personnel. These new communities of users may have significantly different needs to those of the traditional users for whom many existing information services have been designed. For example, administrative users may want to search for and use digital records in quite different ways than researchers might want to access them.

Metadata can also be used to document changing uses of systems and content. That information can then become feed back for future project planning. Well-structured metadata can also facilitate an almost infinite number of ways to search for information, present results, and even manipulate information objects without compromising the integrity of those information objects.

MULTI-VERSIONING

Digital images often exist with multiple versions. These versions may be as simple as creating both a high-resolution “master” copy for archival and research purposes and a low-resolution thumbnail image that can be rapidly transferred over a network for quick reference purposes. Or it may involve creating enhanced images for special purposes.

In either case, there must be metadata to link the multiple versions and capture what is the same and what is different about each version. The metadata should also be able to distinguish what is qualitatively different between the digitized version and the original physical object.

LEGAL ISSUES

Metadata allows repositories to track the many layers of rights and reproduction information that exist for information objects and their multiple versions. Metadata also documents other legal or donor requirements that have been imposed on objects - for example, privacy concerns or proprietary interests. By putting legal information, if it is deemed necessary, into the metadata of a digital record, that metadata will be available for future users if needed.

PRESERVATION

In order for today's digital records to have a chance of surviving migrations through successive generations of computer hardware and software (or removal to entirely new delivery systems), they will need to have metadata that enables them to exist independently of the system that is currently being used to store and retrieve them.

Technical, descriptive, and preservation metadata that documents how a digital information object was created and maintained, how it behaves, and how it relates to other information objects will all be essential.

This descriptive metadata should be migrated along with the digital image so the record will remain accessible and intelligible over time.

SYSTEM IMPROVEMENT AND ECONOMICS

Benchmark technical data, much of which can be collected automatically by a computer, is necessary to evaluate and refine systems in order to make them more effective and efficient from a technical and economic standpoint. The data can also be used in planning for new systems.

INTEROPERABILITY

A variety of mapping rules between different metadata formats are available²⁷, since most metadata standards are open. However, there will still be a certain level of effort required to make sure that images of differing metadata standards can work together.

Generally a single metadata standard should be used for all records in a project, which will facilitate record searching and comparison.

If records are to be imported or exported from or to another project with a different metadata standard, the metadata should generally be converted to a single metadata standard.

13.2 Image Metadata

Metadata or data describing digital images must be associated with each image created, and most of this should be noted at the point of image capture. Image metadata is needed to record information about the scanning process itself, about the storage files that are created, and about the various pieces that might compose a single object.

The number of metadata fields may at first seem daunting. However, high proportions of these fields are likely to be the same for all the images scanned during a particular scanning session.

²⁷ See <http://www.ukoln.ac.uk/metadata/interoperability/> for an extensive list of links to metadata conversion mapping specifications.

For example, metadata about the scanning device, light source, date, etc. is likely to be the same for an entire session. And some metadata, about the different parts of a single physical object (such as the scan of each image in a roll of aerial photos), will be the same for the digital objects that are created. This repetition of metadata will not require entering each individual metadata field for each digital image; instead, these can be handled either through inheritance or by batch-loading of various metadata fields.

14 Image Formats

The following image formats are commonly used by digital image archives, some for archival storage and some for presentation. The file formats use a variety of compression schemes that have comparative advantages and disadvantages, and each serves needs of either bi-tonal, or gray scale, or continuous tone/color images.

14.1 A Note on Editing Lossy Image Formats

All lossy images are subject to further image loss if they are edited in their lossy form. For example, if a JPEG is edited in a JPEG editor, then saved, the image can lose additional information as the compression algorithm is applied again. Additional loss will occur if the image compression ratio is modified, although simply changing the content of the image can cause additional information to be lost. Normally, this is evidenced as blurring of lines and loss of color over time.

The best solution is to create new lossy images from the lossless digital master if it available. If a lossless master is not available, the loss can be minimized in some cases by extracting the lossy image to a lossless format (such as TIFF), and then editing the lossless image before saving it back to lossy form, however, it is still possible to lose information this way.

14.2 TIFF ITU-T.6

A TIFF files use a 24-bit storage format, commonly used by bitmap editors, the TIFF format may be used to store color images. This format is also suited for bi-tonal text documents. This ITU²⁸ specification provides a lossless compression format with good support for color depth, which makes TIFF ITU-T.6 a good format to be used for archival files.

With lossless compression, the data that results from compressing and then uncompressing the image is exactly the same as the original, uncompressed file.

There are multiple versions of the TIFF format, with the ITU-T.6 version being the latest standard. Project staff should make sure this version is used when creating master images.

Some scanning software is able to populate the headers (tags) in TIFF files with information related to the scanner used to create the image, or the scanning process itself. Adobe maintains the TIFF format and assigns tags (both public and private) to companies requesting them. Examples of public tags used by some scanning software are included in Appendix C.4.7 - TIFF header requirements.

Several GIS companies have banded together to support GeoTIFF, an extension to the TIFF format. The GeoTIFF spec defines a set of TIFF tags provided to record geographic coordinates and map projections, and for describing those projections. [43]

14.3 JPEG

JPEG²⁹ is a 24-bit, lossy (some visual information can be lost when images are stored in this format) compression format, which is well suited for screen and print presentation. JPEG is supported by all major computer platforms and by Internet web browsers.

²⁸ International Telecommunications Union., a standards body.

²⁹ Joint Photographic Expert Group. The group responsible for the image format, composed of representatives from the ISO and ITU and other groups.

JPEG provides a high (but lossy) compression ratio for photographic imagery, [20] but is not generally suitable for saving line drawings.

With lossy compression, the picture quality of the compressed file is reduced when compared to the original file, and cannot be restored, except by going back to the original. The advantage of using lossy compression is that the file sizes are much smaller, and image quality is acceptable in most cases. However lossy image formats are not acceptable as archival file formats.

14.4 JFIF (JPEG File Interchange Format)

JFIF (the JPEG interchange format) is a specific implementation of the JPEG standard, commonly used by bitmap editing programs, viewers, and Web browsers. This is the formal name for the JPEG format and is normally only used when there is a need to make a distinction between the JPEG compress algorithm and a JPEG file.

14.5 JPEG 2000

JPEG 2000 is another image format from the JPEG group (ISO-15444). This format uses wavelets for image compression instead of discrete cosine transforms used in traditional JPEGs.

The format can be lossless or lossy, depending on how the image was created. The degree of compression depends on what the “lossy” setting was when the image was created. Variable areas of the image can have different “lossy” tolerances set, adding additional compression gains.

The following figure shows a comparison of a raw rasterization image (24-bit), JPEG200, and traditional JPEG.

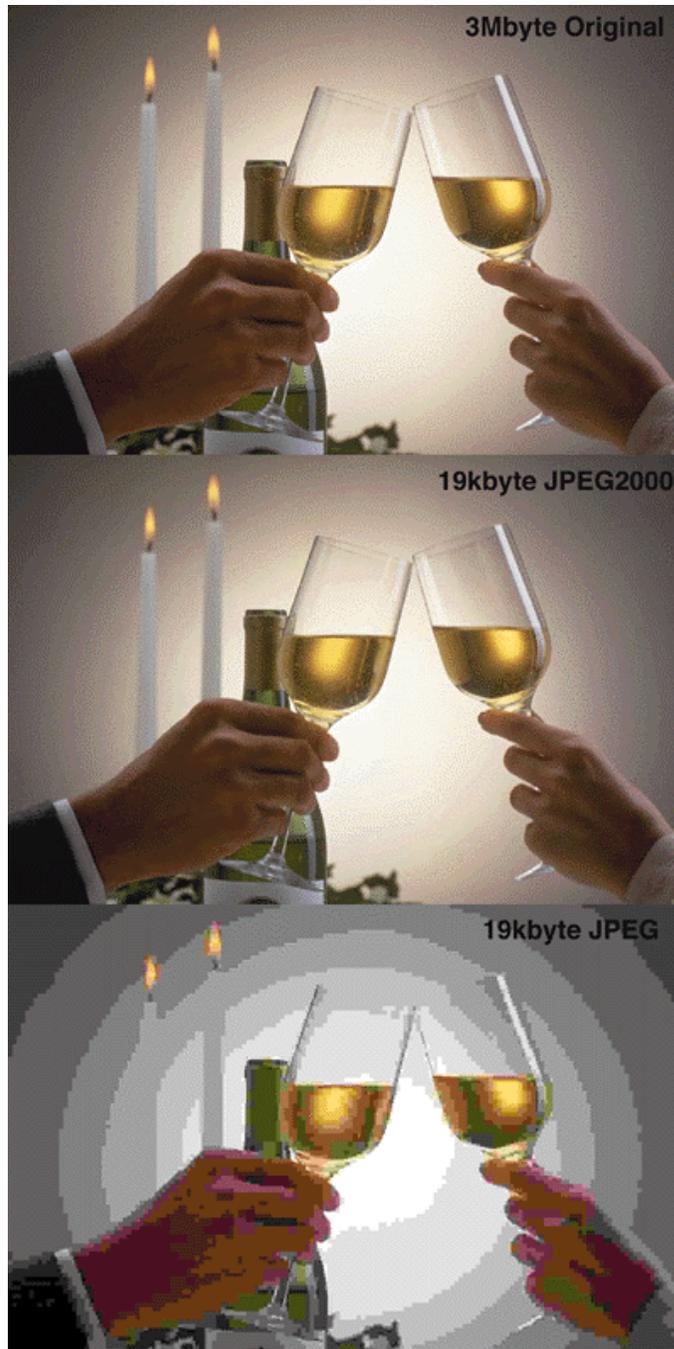


Figure 14-1 Comparison of Raw Image (top), JPEG2000 (middle), and Traditional JPEG (bottom)

Because this file format is not widely used or supported yet, it is not generally recommended for digitization projects. However, it is growing in popularity. Plug-ins exist for many image editing software packages that will add support for this file format. At a future date it may be worth considering lossless JPEG2000 for master digital images and lossy JPEG2000 for thumbnails and derivative images.

14.6 GIF

GIF is an 8-bit, lossless compression format, which is well suited for low-resolution screen display of images. GIF is often used for image thumbnails and screen versions of text documents, and is supported by all major computer platforms and Internet web browsers.

The low color depth (8-bit) is not generally suitable for images with high color depth (e.g. color photographs).

14.7 PNG

The Portable Network Graphic (PNG) format is expected to provide a higher-quality replacement for the GIF, particularly for images delivered to World Wide Web browsers.

Like GIF, PNG is a lossless file format. PNG support is growing, but it is not as popular as the GIF format. PNG compresses image data between 5% and 25% better than GIF.

14.8 PDF

Portable Document Format (PDF) provides a convenient way to view and print images at high resolution, and may also be used to group several files into chapters and books. PDF can provide additional navigational tools such as hyperlinks among pages within a document, and from one PDF document to another. Although this is a proprietary Adobe format, both the file specifications and the viewer software are freely distributed. Plug-ins are available for major web browser to enable them to view PDF files without launching an external viewer.

However, proprietary formats are generally not encouraged for digital masters for a variety of reasons: extra overhead of acquiring plug-ins, lack of interoperability, uncertainty of long-term viability, unknown future support costs, etc.

14.9 CCITT Group 4

CCITT³⁰ Group 4 is optimum for bi-tonal images up to 400 DPI. It was originally intended for encoding fax documents to be sent over ISDN networks. The specialized nature of the format does not lend itself well to mixed document collections, so it is not recommended as a general purpose format for either digital or derived images.

³⁰ Comite Consultatif International Telephonique et Telegraphique, now known as the ITU (International Telecommunications Union). This is an international standards group.

15 Image Format Selection

15.1 Introduction

Imaging is a process by which a physical document (typically paper, film, or drawings) is converted to a computer-readable digital image file. A digital image consists of pixels (picture elements or tonal values in binary code) arranged in columns and rows. The number of pixels per inch determines the image's resolution, which in turn defines the clarity and definition of the image. The height and width of digital images is expressed in pixels, because the size of the image will depend on how it is viewed and the resolution of the monitor that renders it. The resolution of a printed image is measured in dots per inch, which describes how many distinct points of ink a printer can place on paper.

The digital images can be stored on a variety of media. The most common types of storage are magnetic media (e.g. tapes, and magnetic cartridges) or optical media (e.g. CD-ROM and other removable disks). When combined with effective indexing and metadata, digital imaging can provide shorten information retrieval time, allow access to materials for multiple users at various locations, and preserve information that might otherwise be lost due to degradation of the original physical document.

Image files come in many different types of software-dependant formats (e.g. GIF, JPEG, and TIFF). Software-dependant file formats can be read by humans, so computers need software to convert the images back to a human-readable format. Proprietary file formats may not be supported long term by manufacturers and may vary from vendor to vendor. Many file formats use compression to force more data into less storage space and speed image precision, storage, and transmission. Compression may be *lossless* (less compression but no data loss) or *lossy* (better compression, but with some loss of data). Lossy file formats, such as JPEG (.jpg) files, don't necessarily look the same after compression.

If an image contains text, it may be possible to automatically capture the text in addition to the visual image using optical character recognition (OCR) technology. OCR is purely a software feature that reads the image produced during scanning and looks for recognizable patterns (in this case text) in those images. Then, the electronic text can be stored in computer-readable form for search and retrieval purposes. OCR doesn't work for all kinds of documents, particularly for handwritten, poor contrast, unusual type font, or mixed text/image documents. Even when using good quality originals, there will be costs for post-scanning clean up of raw OCR text, since a human user will almost always need to review the interpreted text for errors and make any necessary corrections.

15.2 Image Quality

In almost every case, there is a direct correlation between the production good quality digitized images and the readiness and flexibility with which that image can be migrated across platforms. In order to facilitate a high level of usability, the digitization of physical documents should be done at the highest affordable quality level so that the images produce will be more accessible over the longer term. By maximizing the image quality of the digital master files produced during scanning, managers can ensure the on-going value of their efforts, and ease the process of creating derivative images.

Determination of what the minimum acceptable quality level for digital image should be will depend on the project's planning choices and goals. Project designers need to consider what standard practices they will follow for input resolution and bit depth, layout and cropping, and image capture metrics (including color management). The particular features of the capture device and its software should also be considered.

Benchmarking quality³¹ for any given type of source material can help select appropriate image quality parameters that will ensure the proper amount of information is captured from the source material; they also

³¹ See Anne R. Kenney. Digital Imaging for Libraries and Archives. Cornell University Library, June 1996.

provide a series of test targets and samples that can be compared against the products of day-to-day scanning during quality control.

Digital image collections intended for long term storage and presentation should store from three to four images for each original item: an archival image, derivatives for viewing and a thumbnail for browsing. The master or archival image should capture as much information as possible to preserve the investment in the capture process. Masters should use color rather than grayscale when color is an integral part of the information conveyed by the original object, and any compression applied to the file should be lossless. Grayscale originals (e.g. aerial photographs) should be scanned in grayscale.

Viewing files can be created at any time from the archival image and should be created to provide reasonable access by standard viewers. There are two common viewing images that are normally created: a preview or thumbnail file for the fastest access during initial search and retrieval process and an access image (sometimes called a service or reference image) for more detailed viewing.

While compression is allowed for archival files, some sources [12] discourage its use, as it can pose complications for data migration and raise issues of long-term preservation. When compression is used, it must be lossless and not proprietary.

As the resolution of an image increases, so will its file size. Doubling the resolution (PPI) will quadruple the size of the file. It's clear that the file size can quickly become a limiting factor, which places an upper limit on the amount of information that is saved. The size of a digital image file depends on the size of the original and the resolution of capture (number of pixels per inch in both height and width that are sampled from the original to create the digital image). The file size will also be affected by the number of bytes required to store the color information for each pixel. Typically there are three bytes used for high quality image storage, one for red, one for green, and one for blue (called RGB color).

15.3 Specific Minimum Resolution and File Formats

The intent of the following is to offer guidelines for scanning various types and sizes of original documents, so that the digital master files as captured will record all of the significant visual features in the original item. Capture resolutions are based upon the assumption [11] that a scanning resolution of 600 PPI³² will be sufficient to meet this requirement for most originals in most collections, apart from negatives and transparencies.

In some cases, it may not be possible to create a digital master files which captures all of the visual information present in the original, due to equipment constraints. It is assumed that equipment capable of creating archival quality digital masters would be acquired at some future date, at which point the original digital images would become obsolete and would be replaced by new images scanned at the appropriate resolution. However, projects should make every effort to create digital masters that will be of a sufficient quality so as to server as long-term archival sources.

It is not safe to assume that the original physical object will always be available to create future digital images.

15.4 Reflective Formats

The reflective formats, such as photographic prints, maps, illustrations, and manuscripts, are based on 8.5" x 11" originals scanned at 600 ppi. The 35mm format has a resolution standard of 4200 pixels in the longest dimension, as this is about as much data that most 35mm films can capture. Scanning the 35mm format, which is 1.5" on the longest side, at 2800 ppi will result in compliance with the 4200 pixel standard, or if the same image is printed at 8.5" x 11" scanning at 400 PPI will capture the image with the same resolution.

Aerial photos are generally taken a 300 DPI on a continuous roll of 9" x 9" images. Scanning at a higher image resolution will reduce the error associated with digitization.

³² PPI: Pixels per inch. Similar to DPI (dots per inch) for printed materials.

Note that if you plan to create a film intermediary of the object and then the intermediate image should be of a sufficient resolution so that image data is not lost either in creating the intermediate image, or in creating an intermediate image that will be of too high a resolution to accurately be captured by the scanner.

15.5 Transmissive Formats

Other transmissive formats, such as negatives, slides, and transparencies have a standard of 6000 pixels on the longest side, based on an 8.5" x 11" original, which yields an image with a resolution of just under 600 PPI (6000 pixels / 11").

As will intermediaries created for film, make sure that no information will be lost in the conversion process.

15.6 Oversized Originals – Maps and Engineering Drawings

Due to the fine detail of the graphic and textual elements paper maps and other oversize documents such as engineering drawings present special difficulties for digitization. There can be a huge disparity between the size of the document and the size of the smallest meaningful element that must be made visible online or in printouts. Fine detail requires high resolution scanning, and the result is large file size. File manipulation, storage, delivery, and display all become much more complicated.

Problems users may encounter trying to access and navigate within digital images of large maps include:

- If the image is a high-resolution image (so all of the details are visible) then the image is too big for users to access or manipulate easily.
- When derivatives of the original high-resolution files are provided for access, they are often JPEG versions with considerably reduced resolution. If the resolution is low enough to make files easy to access, the finer details in the images may become illegible.
- Only part of the map image fits on screen at one time. When using the paper document, readers orient themselves to salient features through peripheral vision while focusing closely on details. On screen, it is easy to become disoriented because most of the image is not visible.
- With a paper map, it takes a single glance to follow features such as roads or boundaries from one edge to the other, but on screen it takes continued scrolling. Comparing widely separated details becomes awkward if they are not visible simultaneously.

Scanning oversized documents is difficult since most flatbed scanners are not capable of scanning originals larger than 11" x 17". One scanning alternative is using film intermediaries such as 4x5 transparencies or single-frame microfiche, where the original object fills the body of the microfiche. Thirty-five mm slides are too small to fully capture details on large maps. When originals are not only oversized but also brittle, working from a film intermediary will put less strain on the fragile original. Some loss of quality will result because the film version is one generation removed from the original. However, legible images can be produced from film intermediaries given that the transparency or microfiche is itself carefully made and then scanned with sufficient resolution and appropriate tonality.

Another scanning technique for oversized documents is to scan the document in sections to generate a group of high-resolution files of manageable size. These can either then be joined to form the final archive image or delivered as a set where the entire document is only viewed at low resolution and the user is given the opportunity to zoom into a higher-resolution section of the document.

15.7 Bit Depth and Color Depth

Bit depth is an indication of an image's tonal qualities; specifically it is the number of bits of color data that are stored for each pixel. The greater the bit depth, the greater the number of gray scale or color tones that can be represented. Larger bit depths will also mean larger file sizes.

A bit represents the smallest amount of information that a computer can deal with. It is a single memory location with a value of either one or zero. Bits are commonly grouped in blocks of eight, referred to as a byte. Each bit can have only one of two values; so eight bits can have 2^8 values, or 256 values. Clearly, the more bits (or bytes) that are used to represent an image, the more information it can store. Color (or bit) depth, is usually represented as the number of bits or bytes used to store information about a single pixel (single color block within an image).

Bit depth also has an effect on file size: as bit depth increases, the uncompressed file size increases arithmetically. Scanners may sample at a higher bit depth during scanning, and then produce a final image with a lower bit depth. Sampling at a higher bit depth aids in reducing noise, extends the possible tonal range of the image, and allows the scanner to capture a larger density range without loss of detail.

There are three levels of color depth for scans: bi-tonal, grayscale, and color. The most common bit depths are:

BI-TONAL

Bi-tonal or binary data stores pixel information in a single bit. In this case a pixel is either black (the bit equals zero) or white (the bit equals one). Bi-tonal scanning is best suited to high-contrast documents such as printed text.

When using OCR software, bi-tonal images are generally easier for algorithms to interpret, which can result in better character recognition, although the image may not be as "human readable" as it would be if it were scanned in grayscale or color.

GRAYSCALE

Grayscale is represented using multiple bits per pixel, typically 8-bits, representing shades of gray. Eight bits per pixel means that the pixel will have one of 256 gray tone values. Grayscale is suited to continuous tone documents, such as black and white photographs.

Sometimes, scanning a bi-tonal original in grayscale will produce a more readable image because tones can be used to smooth the transition from black to white and reduce the appearance of stair-stepping.

COLOR

Color is represented using multiple bits per pixel representing color. Color scanning is suited to documents with color information. These three modes of scanning also require some subjective decisions. For example, a black and white typed document may have annotations in red ink. Although bi-tonal scanning is often used for typed documents, scanning in color may be preferable in this case, depending on how the image will be used. Manuscripts, older printed matter, and sheet music may be better served by scanning as continuous tone in grayscale or color to bring out the shade and condition of the paper and the marks inscribed on it.

8-bit color uses 8 bits per pixel, which provides up to 256 color codes. These codes map to a "color palate".

24-bit color is sometimes referred to as RGB color. The color data is split into three 8-bit color channels, representing the color ranges in red, green, and blue (hence RGB). Each color channel can have 256 values, for a total of 16 million different color combinations.

In general, images should be captured at bit depths greater than 24 (which only allows 256 levels for each color channel), standard formats for storing and exchanging higher bit-depth files have not yet evolved, so digitization projects can expect that (at least for the next few years) the majority of digital master files will be 24-bit. Project

planners considering bi-tonal capture should run some samples from their original materials to verify that the information captured is satisfactory; frequently grayscale capture is desirable even for bi-tonal originals.

8-bit color is seldom suitable for digital masters.

15.8 Resolution

The resolution at which you scan is one of the key factors that determine the quality of the images that are produced. Resolution is often expressed as an array: the number of pixels across both dimensions of an image (or more simply as 3000 pixels across the long side), as dpi (dots per inch), or as ppi (pixels per inch).

Higher dpi settings will generally yield a better digital image, because they use more pixels (and therefore, information) in an inch than do the lower dpi settings. However, the higher the dpi, the larger the file size will be.

Scanning at a high resolution is recommended when convert an important collection into digital form to increase access and to build a virtual archive, generate "archival" images, or make prints of the digital image on a good printer. There is a threshold to resolution, however. After a certain point, increasing the resolution of a digital image will not cause visible improvements in the digital image that is produced.

Project planners should be aware of the storage capacity requirements for images produced at the resolution settings required by the project.

15.9 Image Formats for Digital Masters

Digital masters should capture information using color rather than grayscale approaches when color is integral to the information conveyed by the object. Digital masters should never use lossy compression schemes and should be stored in internationally recognized formats.

TIFF is a widely used format, but there are many variations of the TIFF format, and consistency in use of the files by a variety of applications (viewers, printers etc.) is a necessary consideration. In the future, international standardization efforts (such as ISO attempts to define TIFF-IT and SPIFF) may lead vendors to support standards-compliant forms of image storage formats.

Proprietary file formats (such as Kodak's Photo CD or the LZW compression scheme) should be avoided for any long-term project.

Almost all digital archivers recommend that no file compression be used at all for digital master files (although there may be legitimate reasons for considering using a compression format) because of migration issues.

TIFF-6³³ is one of the most popular choices because it is lossless with moderate storage efficiency and has wide support. Limited storage resources may force the issue by requiring the reduced file sizes that file compression affords. Those who choose to go this route should be careful to take into consideration digital longevity issues.

Compression, whether it is lossy or lossless, adds an extra layer of complexity to working with image files. This layer of complexity can complicated migration issues if the compression algorithm can not easily be applied to the files.

Project planners should be aware [12] that some vendors' products may claim to be lossless (primarily those that claim "lossless JPEG"), but will actually be lossy.

³³ The most recent version of the *.TIF file specification. See section 14: Image Formats for more information.

15.10 Image Formats for Derivative Images

Since the purpose of the digital master file is to capture as much information as is possible, as since the file size will increase as the level of information increases, it may be difficult to deliver the resultant images over a network (especially over a slower dial-up connection). Using derivative versions of the master image will probably be needed speed up the transfer process.

Derivative images may also be created to digitally "enhance" the image in one form or another to achieve a particular research goal. Enhancements should be saved in their own files; they should never alter the original digital master file, which should remain unaltered for preservation purposes. Derivative versions may or may not be preserved, depending on the needs of the project.

While it is possible to generate derivative images for network based access "on-demand" from the digital master, the process of generation adds access overhead, and implementing the "on-demand" infrastructure also requires additional resource allocation. While dynamic generation can reduce storage costs (the derived images are can be discarded immediately, then regenerated later) in some way, they can also drive the costs up because the masters must be stored online so that the software used to generate the derivative images can access them as required.

Most digitization projects create the basic derivative images that will be for web-based delivery either in batch or at the time the master image is created. This also adds flexibility in determine what type of access storage should be used for the masters (online, offline, or near-line).

SIZES

Typically two derived images are created. A small preview or "thumbnail" version (usually no more than 200 pixels for the longest dimension) is used for navigation and a larger version that mostly fills the screen of a computer monitor (640 pixels by 480 pixels fills a monitor set at a standard PC VGA resolution), which will be used for general research purposes.

Depending on the need for users to detect detail in an image, a higher resolution version may be required as well.

The full set and sizes of derivative images required will depend upon a variety of factors, including the nature of the material, the likely uses of that material, and delivery system requirements (such as user interface).

Derivative files should be created using software to reduce the resolution of the master image, not by adjusting the physical dimensions (width and height). After reducing the resolution, it may be necessary to sharpen the derivative image to produce an acceptable viewing image (e.g., by using "unsharp mask" in Adobe Photoshop). It is perfectly acceptable to use image processing on derivative images, but this should never be done to masters.

Derivative images will frequently be compressed using lossy compression. Compression algorithms are usually optimized for a particular type of image (e.g. JPEG achieves high compression ratios for pictorial images, but cannot compress images of text very much without introducing compression artifacts), and one should be careful not to use the wrong type of compression scheme for a particular image. Compression algorithms such as JPEG involve a spectrum of options (ranging from high compression ratios that involve visible loss to low compression ratios that involve little visible loss).

Each software implementation of these options labels them differently, and there is currently [12] no objective and interoperable way to declare which of a range of options one has chosen.

ARTIFICIAL V. ENHANCED

Many historical images are faded, yellowed, or otherwise decayed or distorted. Image enhancement techniques can in some cases result in a better digital image, that displays the content of the document more clearly, rather than the decayed condition of the original physical document. If an enhanced image is required, it should be

offered in addition to the digital master, which should depict the condition of the original physical document as closely as possible. By having both images available, users will understand the condition of the original while having a more useable version for general research purposes.

Many software packages offer options that can automatically perform a series of standard operations on images, which can speed the creation of enhanced images. Production of enhanced digital images, however, will most likely not be able to be automated, due to the inability of any one standard transformation procedure to apply equally to all images in a particular project. If automated procedures for image enhancement are not effective, the overhead of creating these images individually will need to be considered by project planners.

COLOR MANAGEMENT

The objective of color management is to control the capture and reproduction of color in such a way that an original print can be scanned, displayed on a computer monitor, and printed, with the least possible change of appearance from the original to the monitor to the printed version. This objective is made difficult by the limits of color reproduction: input devices such as scanners cannot "see" all the colors of human vision, and output devices such as computer monitors and printers have even more limited ranges of colors they can reproduce. In addition to these limits, scanners, printers, and monitors may all be calibrated so as to display (or read, in the case of scanners) the same color differently.

Most commercial color management systems are based on the ICC (International Color Consortium) data interchange standard, and are often integrated with image processing software used in the publishing industry.

Color management systems work by systematically measuring the color properties of digital input devices and of digital output devices, and then applying compensating corrections to the digital file to optimize the output appearance. Although color management systems are widely used in the publishing industry, there is no consensus yet on standards for how (or whether) color management techniques should be applied to digital master files. Though projects may experiment with color management systems for derivative files, until a clear standard emerges it is not recommended that digital master files be routinely processed by color management software.

If color management is used with digital images, the details of how that color management was applied should be recorded in the images metadata. Ideally, the metadata would contain enough information to reverse the color management algorithm and restore the image to its original composition.

Some image processing tools [12] (e.g. Adobe Photoshop 5) default to implementing color management. Users need to beware, and may need to turn off such functions to prevent the digital master (or derived images) from being altered unknowingly.

15.11 Minimum Quality Level – Photographs

These guidelines can be applied to both traditional photographs and aerial photographs. Aerial photographs in the USACE collection typically have an image resolution of 300 DPI, so these standards should be more than sufficient, even if higher resolution images exist.

MASTER IMAGE

Color master images of photographs should be stored in either 8-bit grayscale, if the physical originals were in black-and-white. If the physical originals were in color, they should be stored in at least 24-bit color. Like all master images, they should be stored in an uncompressed format, such as TIFF. The spatial resolution should be 600 to 1000 pixels per inch across the long dimension.

ACCESS IMAGE

Access images should be made with the same tonal depth as the digital masters. That is, 8-bit grayscale for black-and-white originals and 24-bit color images for the access images. Generally, if the master image was made at a higher color resolution (32-bit color), the access image can still be made at a tonal depth of 24-bits. Access images can use compression such as JPEG, with compressions of up to 7:1 - 10:1 for grayscale and 10:1 - 20:1 for color images.

For higher-resolution versions the spatial resolution should range from 200 pixels per inch to 1000 pixels per inch across the long dimension. Image resizing should be based on user requirements, but typical [13] sizes are 640 x 480 pixels, 1024 x 768 pixels, and 1280 x 1024 pixels.

THUMBNAIL IMAGE

Thumbnail images of photographs can be saved in a low resolution, typically 4-bits for grayscale images and 8-bits for storing color. Lossy compression is fine. The compression that is native to the storage format (GIF or JPEG) can be used.

The thumbnail is usually [13] resizes the original to 150 - 200 pixels (or roughly 72 dpi along the long dimension). When there are multiple images with similar composition (for example, images of written documents), a higher resolution thumbnail may be needed so that users can tell the difference between images.

	Master	Access	Thumbnail
Tonal Depth – Color	24-bit	24-bit	8-bit
Tonal Depth - Grayscale	24-bit	8-bit	4-bit
Format	TIFF	JPEG	GIF or JPEG
Compression	Uncompressed	Grayscale = 7:1-10:1 Color = 10:1-20:1	Native to GIF format
Spatial Resolution	600 to 1000 pixels per inch	Resize image: 640 x 480 pixels 1024 x 768 pixels 1280 x 1024 pixels Range from 200 pixels per inch to 1000 pixels per inch for higher-resolution version.	Resize original to 150 - 200 pixels (+/-) across the long dimension 72 dpi

15.12 Minimum Quality Level – Maps and Engineering Drawings

Special rules may apply to determining capture parameters for maps. Color on some printed maps is more important as a coding device than for its precise hue. This may influence the decision when deciding on the tonality (gray-scale or color). It is important to identify the smallest meaningful element (often a thin line) and verify that the resolution is appropriate at this level of detail.

Whereas for historical maps preserving the original integrity of the document may prohibit the use of any image post-processing, for engineering drawings the main purpose in scanning may be to integrate the resulting image with CAD packages. In this case image enhancements (sharpening filters, thresholding techniques) may be useful to obtain the results necessary for input to raster-to-vector transformation software.

MASTER IMAGE

Master images of maps and drawings should be stored in at least 24-bit color, if the originals were grayscale or color, or in 8-bit grayscale, if the physical originals were in black-and-white. They should be stored in either an uncompressed format (e.g. TIFF) or a lossless compression format. The spatial resolution should be at least 600 DPI.

ACCESS IMAGE

Access images should be made with the same tonal depth as the digital masters. That is, 8-bit grayscale for black-and-white originals and 24-bit color images for the access images. Generally, if the master image was made at a higher color resolution (32-bit color), the access image can still be made at a tonal depth of 24-bits. Access images can use compression such as JPEG, with compressions of up to 20:1 depending on the image composition.

The spatial resolution should be at least 1200 pixels across the long dimension for large maps. Small maps may resize the image to 640 x 480 pixels. Higher-resolution versions should range from 1000 pixels to 5000 pixels across the long dimension.

In some cases, this will result in images that can not be displayed in their entirety on a user's monitor. Project planners may want to specify multiple access images with different dimensions, and then let the user choose which one to work with.

THUMBNAIL IMAGE

Thumbnail images of maps can be saved in a low resolution, typically 4-bits for grayscale images and 8-bits for storing color. Compression is allowed, and is usually native to the storage format, typically GIF or JPEG. The thumbnail is usually a resized version of the original, usually at about 150 - 200 pixels. Or roughly 72 dpi along the long dimension

	Master	Access	Thumbnail
Tonal Depth – Color	24-bit	24-bit	8-bit
Tonal Depth - Grayscale	24-bit	8-bit	4-bit
Format	TIFF	JPEG	GIF or JPEG
Compression	Uncompressed	20:1 (depending on image)	Native to GIF format
Spatial Resolution	300 dpi across the long dimension	1200 pixels across the long dimension (large maps) Resize image to 640 x 480 pixels (small maps) Range from 1000 pixels to 5000 pixels across the long dimension for higher-resolution version	Resize original to 150 - 200 pixels (+/-) across the long dimension 72 dpi

15.13 Summary of General Recommendations

Determine the file format that will be used requires that project planners consider at a variety of factors. Before choosing file formats and specifications:

- Define user community, user requirements, uses, and type of material/collection
- Scan at the highest quality you can possibly justify based on potential users/uses/material. Err on the side of quality.
- Do not let today's delivery limitations influence your scanning file sizes; understand the difference between digital masters and derivative files used for delivery
- Many documents which appear to be bi-tonal actually are better represented with grayscale scans; some documents that appear to be grayscale are better represented in color

- Include grayscale target, standard color patches, and ruler in the scan group if reference targets may be needed by some of the users. This can be done simply by making the test targets (See section 10: Sample Standard Tests) available as master images associated with the scan group.
- Use objective measurements to determine scanner settings (do NOT attempt to make the image good on your particular monitor or use image processing to color correct)
- Do not use compression for digital masters, or only use lossless compression (which could complicate long-term preservation)
- Store digital images in a common (standardized) file format
- Define minimum metadata, especially for linking derived images and documenting image enhancements
- Capture metadata in process so that it is available as soon as the images themselves are available
- Capture as much metadata as is reasonably possible (including metadata about the scanning process itself)

16 Storing Images

16.1 Storing Images

Proper storage will help ensure access to and long-term maintenance of image collections. Storage media consists of the materials on which the digital images are written as well as the devices that record, read, and process the information. Choices for the storage of your images will depend on the technical infrastructure you have in place; however, careful consideration of storage choices will help make the investment in image capture and equipment worth the cost, time, and labor.

Digitization projects may need to consider multiple storage media for their digital collection. Planning for adequate backup storage (which may also include offsite storage in case of disaster) is a must.

Other considerations for storage media and systems include:

- capacity of the medium (how much it can store)
- speed (how quickly images can be written, read, retrieved)
- reliability (stability and longevity of the media)
- security (risks of the medium, safeguards built into the medium to protect data)
- scalability (planned growth rates)
- costs (purchase costs, housing costs, training, maintenance, costs of access, cost of migration, etc.)

There are several types of storage media available for online, offline, near-line, and archiving purposes.

16.2 Magnetic disks

Magnetic disks include hard drives and removable or external hard drives. These media are appropriate for online storage of indexing data and access/thumbnail images.

One of the advantages of magnetic disks is that they tend to have very high access speeds compared with other long-term storage media. They also generally have declining costs over time.

Magnetic disks also have limited storage capacity, and can experience rapid technological changes [13], which can mean that they could also rapidly become outdated. They can also be susceptible to electromagnetic degradation.

16.3 CD-ROM

CD-ROMs are most often used for long-term storage of master images or used at stand-alone viewing stations.

CD-ROMs have the advantage of a standard³⁴ for both reading and writing information from and to the CD. They also provide a relatively stable media at a relatively low cost, which makes them well suited to multimedia applications.

On the downside, they have a fixed, limited storage capacity. They are also difficult to network [13], because of their hardware interface. Additionally, there is the question of life expectancy for the hardware and even the ISO standard. In recent years, the DVD has emerged as similar storage media, but it uses different format and hardware specifications.

CD-ROM is a commonly supported format; as such it may represent the lowest common denominator available for sharing data with other groups. Certain groups (e.g. NARA) will only accept digital images on CD-ROM. If

³⁴ ISO 9660

image sharing is a goal of the project, the image sharing media required by other groups should be determined early in the project schedule, so it can be accommodated as part of the normal workflow. [4]

CD-ROMs are also occasionally prone to catastrophic failure (meaning that the disk may crack or shatter, rendering it unreadable), which generally occurs when the CD has some micro-cracking and is loaded into a high-speed CD-ROM drive, where the high rate of rotation causes fracturing.

One of the most common [13] "models" for digital projects is to store master images offline (inactively) on CD-ROM, and to make access and thumbnail versions available--24 hours--online.

Making entire CD libraries available online is not generally feasible, because equipment costs do not scale well. Therefore, master images may need to be requested explicitly, and the CD-ROM containing the image will have to be manually retrieved by project staff.

16.4 Tape

Tape media are most often used to create backups of archival masters.

Like CD-ROMs, tape media are relatively stable (although they can be subject to disintegration [13]) and available for a relatively low cost. Additionally they offer high capacity and easy portability.

Unfortunately, tape media must be stored under proper environmental conditions or the media will disintegrate. An additional disadvantage is that access to the data is relatively slow, because it must be read sequentially from the tape. As such, tapes are generally only appropriate for offline storage.

However, if material stored on tape can be made available online. Additional equipment may be needed to integrate tape storage media into networks. General network maintenance and support will also be necessary for any digital collections that will be accessed via the Web. Staff who are trained in network administration will be an essential part of digital projects and system support.

A restoration plan, and schedule for creating backups will be required for any backup plan, regardless of the storage media used. Whenever possible, industry standards for making backups should be followed.

16.5 Accessibility Levels

ONLINE STORAGE

Online storage refers to media that is access-ready. Retrieval is fast, often in seconds [13]. Online storage requires a reliable medium for accessing information, which will allow multiple users to access information simultaneously.

Any online storage system must consider how authentication and security will be implemented. Limited bandwidth and network/website downtime may also be issues.

NEAR-LINE STORAGE

Near-line storage refers to data that is accessed from a drive. Retrieval is fast, often in seconds. It can be faster or slower than online storage. Retrieval can be slow if multiple users have requested the disc. Near-line storage provides more security and reliability, but has more limited access.

OFFLINE STORAGE

Offline storage refers to data stored “on the shelf”. It must be retrieved by a person. Retrieval time can take minutes to hours. Offline storage provides the benefits of a low cost storage solution with more security and reliability, but at the cost of limited access. Additionally, data stored offline is not easily browsable.

When resource constraints prevent master images from being made available either online or near-line, offline storage is the only available option. To increase availability to users, access images (images with much of the visual information, but stored in a smaller file format) should be made available to user to minimize access time. For many users, the access images should be sufficient for day-to-day usage.

16.6 Access Types

Optical media can be separated into two access categories: read only, and read/write.

READ-ONLY OPTICAL MEDIA

Read-only optical media are optical media that have no recording capabilities, but can provide playback of information that has been mastered onto the media during the manufacturing process. Types of read-only optical media include laser disks, CD-ROMs, and DVDs. Laser disks come in 8" and 12" formats that store still or moving video analog images accompanied stereo audio. Compact disks are generally available only in standard 4.75" formats that store digital audio, video, or text. This format includes CD-ROM and DVD.

READ/WRITE OPTICAL MEDIA

Read/write optical media are optical media that support both the recording and the playback of information. Types of read/write optical media include both re-writable media and WORM³⁵ media. WORM optical media include more traditional CDs, which can only be “burned” once. Re-writable optical media, such as re-writable compact disks (CD-RW), are another option. Refer to the project goals on longevity to determine if there is a need for re-writable media, which can often be more expensive than WORM media.

16.7 Recording Mechanisms

Information is recorded as microscopic areas of reflective differences in a layer attached to the substrate of the optical recording media. These reflections can be detected by a laser, and then "read" by application software. Optical media employ a variety of recording technologies. [4]

16.8 Types of Stability Factors

There are three types [4] of inter-related stability issues associated with optical media: physical stability, data stability, and technology stability.

Of these three inter-related issues, physical stability and data stability are the least problematic. Although there can be failures in data, and degradation of the media, these generally occur with a small frequency. They should still be addressed through periodic maintenance, but physical and data failures only threaten a small portion of a digital collection. Technology stability is a much greater long-term threat to a collection, because it can affect all of the digital records.

In considering optical media, economic issues also apply overall, such as how long the software vendor has been in business, since most optical media is proprietary. Due to their proprietary nature, optical media are also susceptible to sales and marketing decisions, such as where the products are advertised or which integrators work with the products, which should be considered when deciding whether to use optical media, as well.

³⁵ Write-once, read many.

16.9 Physical Stability Factors

The physical stability of optical storage media is directly related to the environment where it is stored. The way in which the media is used will also impact the physical stability of the storage media.

These factors [4] that relate to physical stability have been derived from accelerated aging experiments in which data are continuously written to and read from storage media under various environmental conditions.

Useful media life is determined by measuring the number of errors (also known as "block error rates" or BLERs) for a particular medium over time, below a maximum acceptable level of read/write errors. On the basis of these types of experiments, manufacturers have claimed that the lifespan of optical media ranges from 15 to 200 years. Generally, manufacturers cite longer life spans for recorded versions versus unrecorded (i.e., blank) versions of particular optical media formats. This means that project planners should order media as it is needed; stockpile inventories of unused media should be avoided because deterioration begins at the time of manufacture, not recording.

HUMIDITY AND TEMPERATURE

The amount of water that comes into contact with the media is based on the humidity and temperature of the surrounding environment. Contact with humidity or water can either cause the binding agent (e.g. glue) used in the media to break down or alter the reflectivity of the (semi) metallic coating into which the data are etched. Either situation could affect data retrieval.

While the effect of humidity and temperature conditions varies among types and brands of optical storage media, a generally accepted, recommended [4] temperature/humidity range is 68°F (max. variation $\pm 1^\circ/\text{day}$, $\pm 3^\circ/\text{year}$) and 40% relative humidity (max. variation $\pm 5\%/\text{day}$, $\pm 5\%/\text{year}$).

The manufacturer's environmental storage specifications for specific media should be followed for each storage media. There are currently no national or international temperature or humidity standards for the storage of optical media.

MECHANICAL DEFORMATION

Inappropriate use or storage of optical media can cause warping of an optical disk or scratching of the polycarbonate substrate. Either situation could affect the ability of the laser to read or write information to or from the disk.

In other storage media, physical contact between a reader and the media (e.g. the tape head and magnetic tape) can cause wear and tear on the media. Over time, prolonged access and usage can cause damage to the media and render it unreliable.

DUST AND DIRT

Dust and dirt, whether present during use or in the storage area, can contaminate the media and adversely affect the physical stability of the media or cause read/write errors, which would also affect the data stability of information stored on the disk.

Usually they interfere with the read/write operation between the hardware and media, but they can occasionally cause actual physical degradation of the material.

LIGHT AND MAGNETISM

For re-writable optical media, such as phase-change or magneto-optical storage [4], exposure of the media to light or magnetism could alter the recorded information.

16.10 Data Stability Factors

The stability of information recorded on optical media is closely linked to the physical stability of the media.

Unless the physical stability is ensured, the data stability will be endangered. The only way you can measure physical stability is by measuring errors and monitoring data stability.

Most optical storage media technology [4] has built-in error detection and correction capabilities. The error recovery capabilities and tolerances of various optical media differ. Utilities that monitor the number of BLERs (block error rates) on optical media are used to determine its stability characteristics and migration or recopying intervals.

16.11 Technology Stability Factors

Technological stability of storage media is affected both by the longevity and standardization of hardware and software. When choosing to any storage media or technology, there are a variety of technological factors that should be assessed. Often the longevity and standardization of hardware can be an issue.

Compatibility of a medium across equipment produced by different manufacturers can also present problems, as can the availability of equipment necessary to read a particular medium format.

Adherence of diverse read/write storage media hardware to various computing interface standards (e.g., SCSI³⁶, the ANSI³⁷ family of standards, etc.) that allow your computer to interact with peripheral devices (such as printers, CD-ROM drives, etc.) helps minimize problems, but as standards change, the media will need to change with them.

Longevity/standardization of software factors include persistence of logical file formats (i.e., how long until the file formats become obsolete), and integration/maintenance of the interface between the storage media (and hardware) and various operating systems

The technological stability of storage media could be impacted by technological obsolescence. In addition, hardware and software packages are affected by marketing factors, such as current and anticipated market share or the marketability. The commercial lifespan of hardware can also affect the stable life of any given technology.

This inherent long-term instability of storage technology is best managed through a combination of technology emulation and migration. See section 17: Maintenance and Preservation for more information on long-term preservation.

16.12 Data Recording and Verification

Data files should be inspected to make sure image files open and display properly and that the correct batch has been recorded on the media. Indexing should be verified. If the media is a backup, it should be noted appropriately in the database and metadata.

Indexing will point to the media where the image is stored, but if the physical media can't be found, it won't be of any use. Make sure any removal media (e.g. CD-ROMS, DVD, Tapes, etc.) are clearly labeled. CDs should be labeled on both the case and the disk itself, it is easy for the CD to become separated from its case. If a felt-tip pen is used to label the CD, make sure it is water-based and does not contain alcohol, which can damage the

³⁶ Small Computer Interface System. An interface found in some computers that uses a standard port to provides communication between the computer and its peripherals.

³⁷ American National Standards Institute. ANSI, which is a member of the International Standards Organization, also provides standards for communication between computers and their peripherals.

protective layer of the disc. It is best to write information on the innermost, clear ring. Special adhesive labels are also available for labeling CDs, but the adhesive may have adverse effects on the CD over time. It is best to label only the jewel case or create an insert for it; however

Include information on the removable disk, such as the name of your institution, name of collection, name of project or grant, a unique number for the disc, the beginning and ending file name on the disc, the file formats on the disc, the date the disc was created, the speed, brand, and model of the CD recorder, and relevant scanning information, such as the software used to scan the images, the brand and model of scanner used, and the resolution used to scan the images. At a minimum, the disk should contain enough information to be “refilled” from the indexing database, should it become misfiled.

Removable material, when stored off-line, should be kept in a cool, dry location to minimize degradation.

17 Maintenance and Preservation

There are many advantages that come from digital imaging project. Because high-density storage can be used to make information available to multiple users over the network, there are dramatic improvements in retrieval time. If quality controls are properly put in place, digital images don't lose quality from generation to generation.

However, digital images are not human-readable without computer equipment, which can represent a significant cost, especially when there is a high potential of hardware and software obsolescence. Some sources [5] note that systems change every 18 months to 5 years, software changes every 2-3 years. The useable life expectancy of media is relatively short, even accounting for some backwards compatibility over the course of one or two changes.

“To ensure that media will be readable far into the future, it may be necessary to archive the system along with the media. For a 100-year life, recording systems and sufficient spare parts will need to be archived along with the data storage media. Media with a life expectancy of 20 years are capable of out-surviving existing recording system technologies.”³⁸

If a specific retention period is expected for digital images, the physical media (optical or magnetic media) or directory structure should be the same as other digital images that have the same retention period. This facilitates disposal or migration as a group at future dates, as project needs dictate.

Digital information is notoriously volatile. As time goes on, hardware, software, and storage media become obsolete. Even if the physical medium (e.g., CD, hard drive) that carries the object survives uncorrupted, it is unlikely that a computer will exist that is capable of reading the medium after a number of years. For example, few computers are today are capable of handing 5.25-inch floppy disks, which were popular storage material not long ago (not long ago, relative to the age of most physical collections). For the relatively few computers that might possess the hardware capable of reading the disk, it isn't clear they will have the operating systems or software capable of rendering the machine-readable information into an image that can be viewed with the typical end-users software.

Although no single production decision about format, compression, etc. will guarantee that an object will persist; some decisions are safer than others. Some formats, at least, will be easier to maintain at lower cost across changing technical regimes.

There are no file formats that are guaranteed to be safe, but there are some factors that make some image formats safer than other. If a file format has a known preservation strategy (e.g. as with SGML-encoded ASCII texts where migration through changing regimes is both known and deemed viable and cost effective), that is a positive factor and indication the file format may be a good long-term choice.

Another positive factor is a format that has a good chance of developing a preservation strategy, is to select a format with popular and widespread commercial usage (e.g. PDF, TIFF). Although this doesn't necessarily mean the file format will be easy to migrate, it can be an indication that a migration path may emerge from the commercial sector by the time migration is required.

Based on the historical rate of changes and changes in current technology environment, digitization projects can expect that image files scanned today will eventually need to be rescanned again in the future, either to restore lost digital images or to create digital images with higher resolution specification.

³⁸ Jon van Bogart, NARA 11th Annual Preservation Conference, “Magnetic Tape Storage”, 1996.

Preservation planning can reduce (or eliminate) the costs of rescanning collections by migrating the collection to new systems as technology evolves or emulating obsolete systems so that their data can still be accessed. But, long-term preservation of digital master files requires a strategy of identification, storage, and well as plans for handing migration to new media, including policies about image use and access to them.

Master files, which reflect the most accurate copies of the original physical image, must remain unaltered over time. New derived images can then be created from the master file (which may be better than migrating derived images if a “better” format for the derived images is to be used). Just like the scanning process, migration procedures need quality control procedures to make sure that the digital master and any associated indexing or metadata are not altered in the process.

All critical files (e.g. master files, metadata, and indexing databases) should be backed up and stored in a safe location. If there is corruption or loss of information in the digital records, they can be recovered or restored.

Preservation is a key issue in the digitization field, but there have not yet been any definitive answers to the problem of preserving data in ever-changing technological environment. At the present, there are two general approaches, Emulation and Migration, used to manage preservation.

17.1 Digital Preservation Strategy: Technology Emulation

Emulation assumes that digital records (image, indexing and/or metadata) will be kept in their current encoding format, and that software will be used to emulate the environment used to access the data. This essentially boils down to translating the data, but assumes that some method of accessing the media and hardware will also be available or that the media will be transferred to hardware and storage that can be read from the system running the emulation software.

As an example, consider one of the many emulators that exist for “older” computer systems such as the Commodore PET computer (circa 1982), which originally loaded software programs into memory from using a standard audio cassette. The emulator runs on current hardware and operating systems (e.g. Red Hat Linux, Microsoft Windows XP) mimicking the chipset and instructions of the old hardware, but in order to use old programs for the emulated system, a hardware capable of reading the storage media must exist or the binary data on the media must have been moved to a media that can be accessed by the operating system (e.g. a file on a hard drive or CD-ROM) running the emulation software.

Emulation may be the only strategy available for digital images that were not migrated before their storage format became obsolete.

Emulation assumes that in some cases, it is better (involves less expense and/or less information loss) to emulate on contemporary systems the computer environment in which digital objects were originally created and used. Emulation strategies may be particularly appropriate for complex multimedia objects such as interactive learning modules.

17.2 Digital Preservation Strategy: Technology Migration

Migration is the transfer of digital records (image, metadata, and indexing) from one storage system to another. This migration can occur at all levels, as objects are moved across media as media evolve (e.g. from diskette to CD, and from CD to optical disk or DAT tape). The same is true for the move across software products as the products become outmoded (e.g. from one version of a word-processing or database package to another). The same is true of image formats, as formats evolve (e.g. from GIF to PNG).

Migration requires that digital records be periodically copied from one encoding format to a newer format. The records may also be periodically transferred from one physical media to another in order to avoid deterioration of the tape, disk, or other storage medium.

Clear, consistent metadata is critical to technology migration. When properly implemented, the metadata should detail what file formats are required, and be able to verify that all related files are can also be emulated or migrated. Metadata will also be crucial for future environments, which may include features such as automatic color control, and measurement tools to compare image sizes.

It is not unreasonable to expect vendors to release new versions of software and databases every two to three years (with patches occurring more often). While many vendor offer migration or “upward” compatibility, it can not be assumed that the migration paths will always be available. Digital archives should be prepared for the possibility that their records may not be backwards compatible with new software releases.

Check algorithms should be used, whenever possible, to ensure that no data has been lost in the transfer of digital records. The type of check used will depend on the type of migration.

For example, if transferring a file from one media to another (for example hard drive to CD-ROM) a simple CRC³⁹ check will ensure that all of the information in the digital file was transferred correctly. However, if a format change is required, a more complex method of checking the data conversion, one that will ensure that no pixel information is lost, will be required.

Migration should ideally be lossless. However, when converting from one lossy compression format to another lossy format (e.g. when migrating access images), it is possible that some data will be lost.

17.3 Digital Rosetta Stone

The “Digital Rosetta Stone”⁴⁰ refers to the effort to move away from depending explicitly on one type of hardware and/or software to read digital information. There are several efforts being made along these lines that look at encoding not only the digital image but also information on how to read it using a higher level (and in some cases human readable) format, however no definitive or accepted structure has been created at this point.

The most practical method in place today for preserving digital images is a combination of migrations and emulation.

17.4 Costs

It is difficult to predict just how much a digital imaging project is actually going to cost, and little hard data on the cost, cost effectiveness, and costs over time of digital projects is readily available. Comparing costs to similar past projects can serve as a metric for estimating future costs.

Generally, capture and conversion of data often comprises only 1/3 of the total costs, while cataloging, description, and indexing comprises 2/3 of the total costs. Upfront and ongoing costs can be significant, and economic advantage--and reality--may be better realized through collaborative initiatives or cooperative/regional digitization initiatives, where costs, resources, goals, and expertise can be shared. Initial investment in equipment, staff training, capture and conversion, handling, storing, and housing originals, producing derivative files, CD production, cataloging and building the image database system, and developing Web interfaces are all possible areas of cost for any digitization project.

However, the costs of a project do not end after conversion. Some on going costs that an institution must commit to include the costs of maintaining data and systems over time, including media migration costs and infrastructure costs. Technology obsolescence is inevitable, and the time frame for some components can be

³⁹ Cyclic Redundancy Check. This algorithm calculates a value for each byte of a file based on both the value of the byte and its location in the file. These individual values are combined to create a final “CRC” value for the file as a whole. If a byte is added or removed from the file, a byte is changed, or two bytes are “flipped”, the CRC will change.

⁴⁰ See <http://info.wgbh.org/upf/slides/index.html> for more information.

relatively short (3-5 years), so project planners should anticipate the costs of data migration as a fundamental part of their budgets.

Some surveys [41] of digital archivers report that they undergo media migration roughly every 3-5 years and technology migrations every 4-5 years. Some digital archivists feel that using mainstream technologies with relatively long histories (e.g. Oracle), can help make minimize internal costs of migration because they often offer easier migration paths, although the external (i.e. purchase) costs may not be lower.

17.5 Backups

Regardless of the level of maintenance identified in the project plan, all projects should include a management plan to back-up data files to provide for disaster recovery if the storage media fails or a collection's server crashes. Effective planning should also include resources for storage of back-up files in an off-site location and periodic verification of the integrity of the original and back-up files, as well as a plan for migration to new media.

All digital [41] media degrades over time. Therefore, back ups must periodically be "refreshed" by copying them to new media of the same type. In general, it's better to copy the data to a new physical media rather than reusing old physical media.

It is recommended that industry standards for creating backup be implemented. Data should be copied to either magnetic media (e.g. tapes) or optical media (e.g. CD-ROM). On-site backups should be used for near-term recovery and off-site storage used for long-term recovery.

18 Guidelines for Creating a Request for Proposal

18.1 Overview

Project planning may determine that due to lack of appropriate equipment, lack of staff resources, or large volume of documents to be digitized it is appropriate to outsource the work. The primary benefits of working through vendors could be financial and technical:

- The Corps office does not have to devote space to scanning, nor does it need to convert its space (possibly including construction) to suit electrical and other technical requirements.
- The Corps office does not constantly need to purchase the latest equipment and software. The vendor is responsible for keeping up with the times.
- The Corps office does not have to manage hiring, training of sophisticated specialist skills, and management of staff.
- The vendor copes with costly equipment breakdowns, downtime, and correction of errors.
- The project benefits from the vendor's economies of scale and high productivity.

Even when using the same equipment and methods, some vendors produce a much better product than do others. Identifying and selecting a vendor is not a quick or easy process, especially where large and complex projects are involved. Depending on the project it may not be necessary to go through every step described below, but all projects will need to work through the basics:

- Develop an initial concept of the project and its goals.
- Identify potential vendors.
- Possibly send out an RFI⁴¹ to explain the goals of the project clearly and to discover which vendors are interested and have ideas about how to handle it.
- Establish a project methodology and quality requirements.
- Develop a short list of suggested vendors or resource (BPAs, IDIQs, Interagency contracts, GSA, etc.).
- Write an RFP (request for proposal) and send it to the short list along with samples to be scanned.
- Commerce Business Daily advertisements....
- Communicate with the vendors while they work on their responses, including site visits and meetings when possible.
- Evaluate and compare the vendors' proposals and select the best.
- Write and sign a contract.
- Work with the vendor during the project.

A request for proposal must explain in detail to potential vendors the requirements and specifications for the project, the criteria that will be used to evaluate their proposals, and the specifics of how their bids should be presented.

An outline for an example RFP is contained in Appendix E. This outline is a portion of an RFP used by the Library of Congress and the National Digital Library and should serve as an example of the type and detail of information that should be included in the RFP. Contract information that is not pertinent to the technical issues of the imaging process, workflow, or quality control has been deleted for size/readability reasons. Due to the wide range of document types within the Corps and the individual needs of each office a single sample RFP would not be applicable for every situation. An office can extract from these examples the basic principles needed to construct its own document and can adapt the language to suit its requirements.

⁴¹ Request for Information.

Three complete, detailed sample RFPs from the Library of Congress are available online [23] (one for text documents, one for photographs, and one for film)⁴².

See Appendix C. Example RFP for more information.

18.2 Analyze functional requirements

The first step in preparing the RFP is to have a clear understanding of the goal for digitization and what kind of final product will serve that goal. If the documents to be imaged have historical or preservation status the requirements for digitization will differ from projects whose purpose is to convert drawings or documents into a usable electronic form (current CAD standard, etc.) where the content of the drawing is needed but an exact digital surrogate of the original document is not the intent. Understanding why a project is being undertaken will guide decision making, not only about image quality and user interfaces but also about what work should be accomplished in-house and what work may safely be vended out.

There will always be an in-house component to any digitization operation. The Corps' office that holds the materials to be digitized must take responsibility for:

- Selecting materials to be converted
- Determining the purpose of digitization and the nature of the desired product
- Establishing necessary quality levels
- Verifying the quality of the completed work

The purpose of digitization will greatly affect the choice of vendors. While vendors should be expected to keep up with the latest hardware and software areas and to have fully trained specialist staff, they may do so only in specialized areas. Vendors who specialize in the imaging and rasterization of engineering drawings may not have the experience with preservation of historical maps and photos, and vice versa. Scanning equipment utilized by vendors can range from high-end desktop equipment in the \$1-3,000 range to large commercial drum scanners in the \$50,000-\$100,000 range. Pricing will be commensurate with equipment used as well as manpower needed, so it is important to match the needs of the project with the expertise of the vendor.

It is important to convey the level of quality that is needed to achieve acceptable results, and to do this a request for proposal must state requirements in clear, quantifiable, and verifiable instructions. It should include details for the product in terms of resolution and tonality, file types and storage media, metadata to be recorded, and possibly even the desired type of equipment and software. In the case of deteriorated documents the Corps' project manager must determine if handling of fragile originals off-site is acceptable or if the vendor is required to transport equipment and work on-site. Any requirement to work on-site should be noted in the request for proposal. If it is not clear what sort of product is best it may be desirable to solicit preliminary pilot runs to demonstrate the quality of the vendor's product before the bulk of the digitization occurs.

18.3 Technical requirements

The technical section is the heart of the RFP and should contain all of the information necessary for vendors to write responsive proposals. The technical requirements portion of the RFP should be clear and explicit, with a specific technical description of the deliverables and how compliance will be evaluated. The RFP should be broad enough to allow for different vendors to propose alternative methodologies where appropriate, but specific enough to ensure that they understand the standards they are required to meet. The RFP should be divided into sections that deal with technical requirements, management requirements, pricing, and references.

⁴² Additional resources are available from AIIM [24] and from Cornell University's Department of Preservation and Conservation and the Research Libraries Group [RLG].

After a description of the project in terms of the ultimate objectives, an RFP should include a description of the objects to be scanned in as much detail as possible to help the vendor make an intelligent bid. This would include contents that describe:

- The quantity and physical nature and dimensions of the materials
- A consideration of the varying sizes of the material to be scanned. Are the materials reasonably uniform throughout? If multiple genres are included, describe each group separately.
- A description of proportions. Differing proportions are difficult to work with as are unusually large or small materials.
- A consideration of the content of materials. Is there an intellectual structure to the materials that must be maintained (as in multi-page documents)? Will vendors be able to batch each type of material, or must materials be handled in an order that mixes different sizes and types?

Scanning instructions should include:

- Detailed instructions about the preferred methodology, including resolution, tonality, bit depth, file formats, compression, platform, and storage media
- Instructions for producing derivative images as well as the master images
- Definition of the required level of accuracy and how the institution will evaluate it

Storage and indexing requirements should include:

- Instructions on file naming and metadata
- How to format file names
- Whether pre-existing identification numbers or other information must be keyed in
- What information about processing and equipment must be reported (*e.g.*, kind of scanner used, its settings, color definitions used, date of capture, description of film stock if film intermediaries are scanned)
- Requirements for how the data should be coded in and laid out, and what
- Type of spreadsheet or database to use

18.4 Project management requirements

Project management requirements should include information about:

- Schedule, including weekly/monthly deliveries, deadlines, and turnaround time
- Document receipt and sorting instructions
- Handling requirements, including lighting levels if that is an issue
- Shipping requirements for original materials
- Specify how errors will be defined and corrected, what error correction will be cost-free, and what will be additional charges
- Specify insurance, security, and storage environment while materials are at vendor and in transit
- Specify frequency of reports and invoices and what information they must contain
- Specify deadlines and penalties for missed deadlines

18.5 Vendor response requirements

The RFP should ask vendors to:

- List the hardware and software they would use, including storage media
- Recommendations for resolution, tonality, file formats, compression and so forth if not specified for a specific document type
- Specify their quality control procedures

- Describe their production capacity and document that they can accomplish the work at the specified quality within the timeframe
- Explain how delivery of materials and files will be accomplished (vendor pickup, courier, UPS, or other)
- Describe environmental controls in the facility if that is an issue for original materials.
- Provide the name and qualifications of the project manager
- Supply references for similar work done with other libraries, archives, or museums
- Scan a representative sample that represents a fair cross section of the materials, including both easy and difficult items (If the originals are valuable or fragile, the sample should consist of reasonably similar items that are less valuable or are expendable.)
- Respond with a price proposal
- State prices in specified units of measure, for instance per page, per image, or whatever is appropriate
- Include costs for data input, cost of storage media, shipping, insurance, and any other additional charges
- Determine if prices are firm for the duration of the project
- Provide suggestions for alternative methods that can accomplish the project at the same level of quality.

18.6 Quality control requirements

A request for proposal also should detail the quality control actions of both the vendor and the Corps' project personnel. Careful quality control assurance by the Corps' representative should provide means to recognize whether the product returned by the vendor is what was requested and the steps that will be taken if it is not. No one should take for granted that work is being done as specified without verification through detailed inspection of the vendor images and files. A RFP should specify terms for accepting the product by defining the minimum acceptable level of accuracy

Previous sections described how to carry out quality control on images. Accuracy of vendor-supplied file names and metadata also must be verified, since erroneous metadata or keyed file names in essence mean that an image is lost. Calling up image after image and examining them carefully for flaws is a very time-consuming operation, but it is essential to ensuring that the vendor's product meets specifications. When working with an unfamiliar vendor, it is especially important to carry out thorough quality control as early as possible in the project. Timeliness in returning errors is important, since:

- Tracking error rates helps determine trends needing correction
- It prevents the vendor from continuing to replicate errors
- It alerts the vendor to problem procedures or ill-trained staff
- There is normally a cut-off date (often several months) after which the vendor will no longer accept errors for free correction.

If the project is small it may be possible to examine every image, but most projects are too large. In this case, recommended procedures are as follows.

- Set up a manageable first shipment that will be due back at an early date in the project.
- Perform careful quality control on 100% of the images and metadata in the first returned shipment.
- Record all errors in detail on quality control worksheets.
- Evaluate the errors to determine which are the institution's responsibility due to flaws in the institution's own procedures, to information that the institution failed to give the vendor, or to variations in the materials being scanned that the vendor should have been warned of. The institution will need to pay to have these errors corrected and obviously will need to revise procedures to avoid them in future.
- Determine which errors are due to mistakes by the vendor. If the percent of errors is higher than the agreed-upon rate, return the entire shipment with a full explanation of the errors and require the vendor to start over and produce a new batch. As necessary, discuss changes in vendor procedures for image capture and quality control.
- Repeat the 100% inspection. If the error rate is still too high, send it back again to be redone.

- If the vendor's work meets or is lower than the agreed-upon error rate, return only the individual problem cases for corrections.
- Continue 100% inspection for the first two to three shipments.
- Once it is clear that the vendor is regularly returning a product below the acceptable error rate, cut back to a lower percent (often 10%) inspection of every shipment. This does mean that a few errors will go undetected until the day some user tries to access those images.
- If the error rate begins to climb, return to 100% inspection until the problem is identified and solved.

Appendix A. References

The contents of this report are based on the information located in the sources listed below.

- [1] NARA's Strategic Directions for Federal Records Management
July 31, 2003 http://www.archives.gov/records_management/initiatives/strategic_directions.html
- [2] NARA's Scanning standards http://www.archives.gov/research_room/obtain_copies/scanning.html
- [3] Hyperdictionary, technical definitions, <http://www.hyperdictionary.com>
- [4] NARA's Optical Media FAQ
http://www.archives.gov/records_management/policy_and_guidance/frequently_asked_questions_optical.html
- [5] NARA's Imaging Records FAQ, updated 4/23/2003
http://www.archives.gov/records_management/policy_and_guidance/frequently_asked_questions_imaged.html
- [6] Appendix G: Scanner Survey Questionnaire
<http://memory.loc.gov/ammem/pictel/appg.htm>
- [7] Manuscript Digitization Demonstration Project: Project Finding
<http://memory.loc.gov/ammem/pictel/mddp3tc.htm>
- [8] Project overview, CADD/GIS.
http://tsc.wes.army.mil/projects/projects.asp?fy=03&prj_id=2003.036
- [9] Arizona State Library, Archives and Public Records, Digital Imaging Task Force, *Digital Project Guidelines*, March 2000, version 1.3, <http://www.lib.az.us/digital/>
- [10] Metadata Encoding and Transmission Standard, <http://www.loc.gov/standards/mets/>
- [11] California Digital Library, Digital Image Format Standards, July 9, 2001,
<http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf>
- [12] California Digital Library, Best Practices for Image Capture, Version 1.0, February 2001,
<http://www.cdlib.org/news/pdf/BestPracticeImageCapture.pdf>
- [13] Colorado Digitization Program, General Guidelines for Scanning,
http://www.cdpheritage.org/resource/scanning/std_scanning.htm
- [14] Western States Digital Standards Group, Digital Imaging Work Group, Western States Digital Imaging Best Practices, Version 1.0, January 2003, http://www.cdpheritage.org/westerntrails/wt_bpsscanning.html
- [15] Getty Standards Program, Getty Research Institute, Introduction to Metadata: Pathways to Digital Information, (2000) http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html
- [16] Dublin Core, <http://purl.org/DC/>
- [17] Institutes of Museum and Library Services, A Framework of Guidance for Building Good Digital Collections, (November, 6, 2001), <http://www.imls.gov/pubs/forumframework.htm>
- [18] Best Practices for Image Capture, Version 1.0 February 2001
Maintained by the CDL Technical Architecture and Standards Workgroup
- [19] Image resolution of the human eye, <http://white.stanford.edu/html/numbers/node1.html>
- [20] "Searching for trouble free Document Image Processing Solutions: Uncover Document Image Metrics That Point to Best-Fit Technologies for No-Fault Imaging and Workflow", AIIM International, Lee A. Freidman, 2000
- [21] "Image Digitization - Quality Standards for Digital Images and Digital Imaging Systems", Library Preservation at Harvard, <http://preserve.harvard.edu/resources/digimagequality.html#reports>
- [22] "Recommended Practice for Quality Control of Image Scanners", ANSI/AIIM MS44-1988, December 1993.
- [23] Library of Congress, Building Digital Collections, Technical Information and Background Papers,
<http://memory.loc.gov/ammem/ftpfiles.html>
- [24] "Electronic Imaging Request for Proposal Guidelines, ANSI/AIIM TR27-1996".
- [25] RLG, <http://lyra.rlg.org/preserv/RLGtools.html>
- [26] Federal Geographic Data Committee, Content Standard for Digital Geospatial Metadata,
<http://www.fgdc.gov/index.html>
- [27] "Some Thoughts about Image Quality", R. T. Moore, <http://members.aol.com/dbarnesphd/imgquty.htm>
- [28] "Moiré Pattern", <http://mathworld.wolfram.com/MoirePattern.html>
- [29] "Moving Theory into Practice – Digital Imaging Tutorial", Cornell University Library/Research Department, 2003
- [30] "Geospatial Metadata Standards", http://www.fgdc.gov/metadata/meta_stand.html, April 16, 2003.

- [31] "Spatial Data Standard for Facilities, Infrastructure, and Environment (SDSFIE)", <http://tsc.wes.army.mil/products/tssds-tsfmts/tssds/html/>
- [32] "MARC 21 Metadata Specifications", Library of Congress – Network Development and MARC Standards Office, September 16, 2003, <http://www.loc.gov/marc/>
- [33] "Text Encoding Initiative", September 6, 2003, <http://www.tei-c.org/>
- [34] "Imaging Systems: The Range of Factors Affecting Image Quality", Digital Library Federation, Donald D'Amato, 2000 <http://www.rlg.org/visguides/visguide3.html>
- [35] "Using Dublin Core", Dublin Core Metadata Initiative, August 26, 2003, Diane Hillmann, <http://dublincore.org/documents/usageguide/>
- [36] "About the ULAN", Getty University, 2000, http://www.getty.edu/research/conducting_research/vocabularies/ulan/about.html
- [37] "About AAT", OCLC, June 3, 2003, <http://www.oclc.org/news/announcements/announcement21.htm>
- [38] "Quality Assurance", TASI, 2003, <http://www.tasi.ac.uk/advice/creating/quality.html>
- [39] "Digital Best Practices", Washington State Library, <http://digitalwa.statelibrary.gov/newsite/scanning.htm>
- [40] "JPEG2000 Wavelet Compression Spec Approved", R. Colin Johnson, EE Times, Dec. 29, 1999 <http://www.eetimes.com/story/OEG19991228S0028>
- [41] "Digital Archive, Main Report", ICSTI, 1999, www.icsti.org/99ga/digarch99_MainP.pdf
- [42] "Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative – Final Report", Contract Number: N66032-94-D-0012, Intergraph Solutions Group, http://tsc.wes.army.mil/downloads/ADCS_Final_Report_Main.pdf
- [43] "GeoTIFF - A standard image file format for GIS applications" <http://www.gisdevelopment.net/technology/ip/mi03117abs.htm>

OBTAINING TARGETS:

Color and grey scale targets are manufactured by Kodak, Fuji and AGFA and are available from professional imaging equipment dealers. They are also available through the following Web sites:

Kodak

<http://www.kodak.com/cgi-bin/webCatalog.pl?product=KODAK+PROFESSIONAL+Q60+Targets&cc=US&Ic=en>

Monaco Systems

<http://www.monacosys.com/targets.html>

Filmtools

<http://shop.store.yahoo.com/cinemasupplies/maccol.html>

Coloraid

<http://www.targets.coloraid.de/>

The Graphic Arts Technical Foundation (GATF) supplies various devices and many targets for color management. Aimed primarily at the print industry, these include target images (including the GretagMacbeth ColorChecker) suitable for subjective scanner calibration as well as many output tests used to calibrate digital output and commercial print. Further details are available at:

<http://www.gain.org/servlet/gateway/publications/processcontrols.html>

Resolution targets are available from:

Sine Patterns

<http://www.sinpatterns.com>

Edmund Industrial Optics

<http://www.edmundoptics.com/IOD/DisplayProduct.cfm?productid=1790#Technical%20Images>

Appendix B. Definitions

Abbreviation / Reference	Definition
AACR2	Anglo-American Cataloging Rules, Second Edition
AAT	Art and Architecture Thesaurus. One of several popular standard vocabulary lists focused on art, architecture, and material information.
Aliasing	Group of image defects generally caused by differences in resolution between the original source and scanner.
ARIMS	Army Records Information Management System. A record management system available to the USACE that provides an indexing and metadata database, as well as access control to digital records.
CADD	Computer-Aided Design and Drafting
CCD	Charge-coupled device. A type of optical detector used in scanners.
CCITT	Standards group, now known as the ITU
CELIO	Corp. of Engineers Library Information Online http://www.usace.army.mil/library/catalog.html
CMS	Color Management System
DPI	Dots per inch
DTD	Document Type Definition
IEEE	Institute of Electrical and Electronics Engineers
EAD	Encoded Archival Description
FGDC	Federal Geographic Data Committee, a metadata standards group.
GIS	Geographic Information System
ICC	International Color Consortium
ISO	International Standards Organization.
ITU	International Telecommunications Union, a standards body
LPI	Lines per inch
MARC	Machine-Readable Catalogue
MARKS	Modern Army Record Keeping System. An older Army record management system that has been replaced by ARIMS.
METS	Metadata Encoding Transmission Standard
Moiré	Image defect resulting from pattern interference, generally manifested as a beat frequency.
NARA	U.S. National Archives and Records Administration
PMT	Photomultiplier tube. A type of optical detector used in scanners.
PPI	Pixels per inch
PURL	Persistent URL
RFP	Request for Proposal
SDSFIE	Spatial Data Standards for Facilities, Infrastructure, and Environment. A CADD/GIS standards group for geographic metadata.

SGML	Standard Generalized Markup Language
SPI	Samples per Inch
TASI	Technical Advisory Service for Images
ULAN	Universal List of Artists Names. A standard, collaborative vocabulary list of material creators and authors.
URL	Universal Resource Locator

Appendix C. Example RFP

**SAMPLE ONLY – NOT TO BE USED AS A
TEMPLATE**

Solicitation No. _____

Contract No. _____

(Excerpts Taken from Library of Congress Requests Proposals For Digital Images of Pictorial Materials)

TABLE OF CONTENTS

**Contracts
AND
Logistics
Services**

PREPROPOSAL CONFERENCE

**Date:
Time:
Place:**

Requests for Proposals

Table of Contents

SECTION A Solicitation, offer and award.

PART I - THE SCHEDULE

SECTION B Supplies or services and prices/costs.

Lot 1 - Mandatory Images -- Base Year
Lot 2 - Mandatory Images -- Base Year

SECTION C Description/specification/work statement.

C.1 - BACKGROUND
C.2 - SCOPE OF WORK
C.3 - LIBRARY FURNISHED MATERIALS AND FACILITIES
C.4 - GENERAL IMAGING REQUIREMENTS
C.5 - GENERAL NAMING REQUIREMENTS FOR FILENAMES, DIRECTORIES, AND ASSOCIATED DATABASES
C.6 - GENERAL HANDLING AND SCANNING REQUIREMENTS
C.7 - WORKFLOW AND PROJECT MANAGEMENT
C.8 - RELATED SERVICES
C.9 - DELIVERABLES AND DELIVERY

SECTION D Packaging and marking.

SECTION E Inspection and acceptance.

SECTION F Deliveries or performance.

SECTION G Contract administration data.

SECTION H Special contract requirements.

PART II – CONTRACT CLAUSES

SECTION I Contract clauses

PART III – COLLECTION SPECIFIC REQUIREMENTS

SECTION J

Lot 1 – Historical American Buildings Survey/Historical American Engineering Record Requirements

Lot 2 – Civil War Map Fiche Group Requirements

PART IV - REPRESENTATION AND INSTRUCTIONS

SECTION K Representation, certifications, and other statements of offerors.

SECTION L Instructions, conditions, and notices to offerors.

SECTION M Evaluation factors for award.

RFP XXX

Section A

(include standard form 33 – Solicitation, Offer and Award)

RFP XXX

Section B

(specify types of materials, size, quantity)

CONTINUATION SHEET				
CLIN	LOT 1 - MANDATORY IMAGES -- BASE YEAR	QTY	UNIT PRICE	AMOUNT
B.1.1	IMAGE SET: 5UA, CRI, THM			
	Transmitted-light Materials			
01	B&W Negative, 8 x 10	9,000		
02	B&W Negative, 5 x 7	21,639		
03	B&W Negative, 4 x 5 or smaller	44,220		
04	Color Transparency, 4 x 5 or 5 x 7	50		
	Reflected-light Materials			
05	Cardboard mounted and unmounted prints in mixed sizes from 3 x 5 through 11 x 14, including crop to individual pictures on multiple picture mount, mixed b&w and color	7,500		
B.2.1	PROGRAMMING AND PROCESSING ACTIVITIES (fully-loaded hourly labor)			
01	Technician	150		
02	Programmer	100		
03	Project Leader	100		
B.5.1	STARTUP AND TESTING ACTIVITY	1	LOT	
	TOTAL BASE YEAR COST - MANDATORY IMAGES - LOT 1			

CONTINUATION SHEET				
CLIN	LOT 2 - MANDATORY IMAGES -- BASE YEAR	QTY	UNIT PRICE	AMOUNT
B.6.1	IMAGE SET: CUA, THM (OPTIONAL REQUIREMENT)			
01	Maps from color 105mm microfiche--initial segment	100		
02	Maps from color 105mm microfiche--additional segment for concatenation to create single map image	750		
B.7.1	PROGRAMMING AND PROCESSING ACTIVITIES (fully-loaded hourly labor rates)			
01	Technician	50		
02	Programmer	100		
03	Project Leader	100		
B.8.1	LOT 2 - STARTUP AND TESTING ACTIVITY	1	LOT	
	TOTAL BASE YEAR COST - MANDATORY IMAGES- LOT 2			

RFP XXX

Section C

C.1 BACKGROUND

C.1.1 The Library of Congress and the National Digital Library Program

C.2 SCOPE OF WORK

The Library requires the creation of digital images of ... *(maps, drawings, and photographic collections. Describe use and purpose of document collection.)*

The delivered sets of images shall be coherently and logically named and/or numbered and shall be placed in delivery directories with prescribed characteristics. The vendor will also produce associated metadata. *(If vendor is to convert to other document formats, ie text (via OCR) or vectorization add description.)*

C.2.1 Workflow

Work shall be performed under either LOT 1 or LOT 2 requirements which may be awarded as one (1) or two (2) separate contracts. LOTS 1 and 2 are differentiated by image capture site. Many of the materials which shall be digitized are processed and currently available to the public. To minimize the removal of materials from use by patrons of the Library, the image capture of most collections shall be carried out on the premises of the three Library of Congress buildings on Capitol Hill under LOT 1. Work under LOT 2 shall include historical maps scanned from microfiche. The LOT 2 scanning shall be performed offsite at the contractor's facility.

The delivery schedule and workflow requirements are provided in Section F.

C.2.2 Startup and Testing Phase

Work under both LOT 1 and LOT 2 shall begin with startup and testing activities designed to resolve various technical details and to confirm and finalize the definition of technical elements.

C.2.3 Representative Collections

The materials which shall be digitized vary. The following list enumerates a number of representative collections for which production planning is under way. The Library may add or substitute collections with similar technical requirements. Descriptions and specifications of some of the actual materials are included in the referenced Section J Attachments.

C.2.3.1 From the Prints & Photographs Division

Historic American Buildings Survey/Historic American Engineering Record (HABS/HAER) (80,000 items)--(Section J, Attachment 1)

C.2.3.6 From the Geography and Map Division

Civil War maps on microfiche (ca. 1,000 transmitted-light items)--(Section J, Attachment 6)

C.3 LIBRARY FURNISHED MATERIALS AND FACILITIES

The Library will furnish to the contractor the original materials to be scanned. Items in LOT 1 will be brought to the designated workspace(s) in the Library of Congress buildings (Madison and Adams) located on Capitol Hill. Items in LOT 2 will be packed for pickup/shipments to the contractor's facility.

The following terms describe portions or segments of materials to be digitized:

collection

A coherent group of materials as held by the Library and typically described as a unit in cataloging or finding aids. Example: the 7,000-picture Civil War photograph *collection* which occupies 16 vertical-file drawers.

batch

A block of materials within a collection that shall be treated as a production unit, i.e., task order and delivery requirements will be stated in terms of batches. Example: each four-drawer Civil War photograph file cabinet contains from 1,500-2,000 pictures which constitutes a *batch* for that collection.

scan group

A block of materials of a convenient size for Library personnel to deliver to the contractor and sufficiently large enough to provide work for one or two scanning sessions. Example: each drawer of Civil War photographs containing from 400-500 pictures constitutes a *scan group* for that collection.

C.3.1 LOT 1 - Work Space

For the scanning of the materials in LOT 1, the Library will provide work space, access to electrical outlets, and telephones.

C.3.2 LOT 1 - Original Items

LOT 1 materials shall be captured in *scan groups*. The items to be delivered to the contractor may be stored individually in various types of containers and also be stored as a group in various file drawers. The storage medium will vary from collection to collection.

The size of the scan group will depend upon the value of the original, its fragility, the availability of secure storage at the scanning site, and other factors. At a minimum, enough material for each day's scanning session will be provided.

For most collections of housed items, e.g., negatives in envelopes or prints in mylar, all handling and scanning labor shall be performed by contractor personnel which shall include removing items from storage containers one at a time, performing the scanning and associated record keeping, and replacing the items in the housing after capture has been completed. In some situations, such as with fragile or rare items, a Library technician will assist the operator by removing the item from the housing, handling the materials to be scanned at the time of image capture, and subsequently rehousing the original object. Task order instructions will be provided for specific collections.

Each original item furnished will be marked with a *physical identification number*. Prints typically have an identification number written on the back. Negatives typically have the negative number (the physical identification number for a negative) written on the storage envelope.

C.3.3 LOT 2 - Original Materials

C.3.3.2 Civil War Microfiche

The Library will prepare and pack the fiche to be digitized for shipment to the contractor in batches and will notify the contractor's approved carrier when the shipment are ready. Each batch will be accompanied by an inventory that lists each original fiche. The contractor shall validate each shipment upon receipt and return a Receipt of Materials Form (see Section J, Attachment 7). A written record shall be made of any other deviations, and the COTR shall be notified of such within 24 hrs. To minimize the potential for loss or damage, the Library shall ship fiches in batches of about 500. It is anticipated that no more than 1,500 fiche shall be on the contractor's premises at any one time. This will permit one incoming, one batch in production, and one batch in preparation for return. If a completed batch of digital images (shipment of 500) is rejected for any reason, the Library will halt further shipments to the contractor until corrections are made and the batch has been accepted.

C.4 GENERAL IMAGING REQUIREMENTS

The contractor shall produce sets of digital images following both the general requirements included in Section C and also the collection specific requirements included in the Section J, Attachments. All general requirements shall apply to both LOT 1 and LOT 2 unless indicated otherwise. Additional collection specific requirements not included will be provided as statements of work prior to issuance of task orders.

C.4.1 Mandatory Image Types--Format, Resolution, and Compression

With the exception of the cartographic images, the Library's mandatory requirement is that each original item shall be reproduced as a set of three digital images: (1) an uncompressed archival image (either 3000, 4000 or 5000 pixels--the specific type to be produced for a given collection will be stated in the task order), (2) a compressed reference image, and (3) a thumbnail image (See C.4.3.1, C.4.3.1.1, C.4.3.2, and C.4.3.3). The requirement for cartographic images is that each original item be produced as an archival image as specified in C.4.3.5 and a thumbnail image as specified in C.4.3.3.

C.4.3 Image Format, Resolution, and Compression

Image requirements in terms of format, resolution, and compression for both the mandatory and desirable images are as specified below.

C.4.3.1 5UA, 4UA, or 3UA: 5000-, 4000-, or 3000-PIXEL UNCOMPRESSED ARCHIVAL IMAGES (Mandatory)

- Spatial resolution of approximately either 5,000 (5UA), 4000 (4UA), or 3,000 (3UA) pixels as specified on the image's long side--with the short side falling where it may. This resolution shall be the actual optical resolution of the capture (or a reduction therefrom) and shall not represent values achieved by interpolation.
- No sharpening or other enhancement
- Uncompressed
- "Intel" TIFF, with ver. 5.0 or 6.0 headers (content specified elsewhere in this Statement of Work)
- Must work in IBM-compatible environment

C.4.3.1.1 Mandatory Tonicity

- Tonal (pixel-depth resolution): Color: 24 bits-per-pixel; black and white: 8 bits-per-pixel

C.4.3.1.2 Desirable Tonicity

- Tonal (pixel-depth resolution): Color: 36 or more bits-per-pixel; black and white: 12 or more bits-per-pixel

C.4.3.2 CRI: COMPRESSED REFERENCE IMAGES (Mandatory)

- Spatial resolution approximately 640 pixels on the image's long side--with the short side falling where it may. This resolution shall be the actual optical resolution of the capture (or a reduction therefrom) and shall not represent values achieved by interpolation.
- Tonal (pixel-depth resolution): Color: 24 bits-per-pixel; black and white: 8 bits-per-pixel
- Sharpen or other enhancement shall be at a level as defined and determined during the contract startup and test activity
- JPEG compression to yield average compression of 10:1 for grayscale and 15:1 for color
- JFIF format/headers
- Must work in IBM-compatible environment

C.4.3.3 THM: THUMBNAIL IMAGES (Mandatory)

- Spatial resolution approximately 150 pixels on the image's long side--with the short side falling where it may.
- Tonal (pixel-depth resolution): 8 bits-per-pixel
- Palettes optimized (adaptive palettes) for each image

- Sharpen or other enhancement shall be at a level as defined and determined during the contract startup and test activity
- Uncompressed
- "Intel" TIFF, with ver. 5.0 or 6.0 headers (content specified elsewhere in this Statement of Work)
- Must work in IBM-compatible environment

C.4.3.4 5DI, 4DI, or 3DI: 5000-, 4000-, or 3000-PIXEL COMPRESSED DISPLAY IMAGES (Desirable)

- Spatial resolution of approximately of either 5,000 (5DI), 4000 (4DI), or 3,000 (3DI) pixels as specified on the image's long side--with the short side falling where it may. This resolution shall be the actual optical resolution of the capture (or a reduction therefrom) and shall not represent values achieved by interpolation.
- Tonal (pixel-depth resolution): Color: 24 bits-per-pixel; black and white: 8 bits-per-pixel
- Sharpen or other enhancement shall be at a level as defined and determined during the contract startup and test activity
- JPEG compression to yield average compression of 10:1 for grayscale and 15:1 for color
- JFIF format/headers
- Must work in IBM-compatible environment

C.4.3.5 CUA: CARTOGRAPHIC UNCOMPRESSED ARCHIVAL IMAGES (OPTION)

- Spatial resolution of 300 dpi as measured against the original paper map that has been copied onto a 105mm fiche (see Section J, Attachment 6). The scanning resolution applied to the fiche shall be high enough to yield 300 dpi when the fiche reduction ratio is considered and shall be the actual optical resolution of the capture and shall not represent values achieved by interpolation. Image-tile concatenation may be required to produce these images.
- Sharpen or other enhancement shall be at a level as defined and determined during the setup activity for this collection
- Uncompressed
- "Intel" TIFF, with ver. 5.0 or 6.0 headers (content specified elsewhere in this Statement of Work)

- Must work in IBM-compatible environment
- **Tonal (pixel-depth resolution): Color: 24 bits-per-pixel; black and white: 8 bits-per-pixel**

C.4.4 Image Tonal Range

The tonal range of the delivered digital images delivered shall be representative of the original scene or artifact or, in the case of images whose source is a photographic negative, of the expected representation of the original scene when the negative is reproduced as a positive print.

- For prints (black-and-white or color), the objective is to reproduce the items as they exist in the collection.
- For negatives and original positive transparencies, the objective is to create a positive image in a manner that may be compared to creating a print (black-and-white or color) in a darkroom.

Providing images with acceptable tonal qualities shall require that the scanning operator exercise judgement when producing the images. The operator's judgement required to achieve the required outcome shall be especially critical when imaging color items. Utilization of general imaging industry standards and those as agreed to and established during the contract startup and testing phase shall be followed. Additionally, consultation

with Library staff may, when necessary, be required in order to ensure that appropriate operator judgements are made throughout the run of a particular batch or scan group.

C.4.4.1 Tonal Value (Mandatory)

Tonal value for all mandatory images and the desirable compressed display image shall be as follows:

- For a black-and-white scene with a typical range of brightness, e.g., a landscape in daylight or a conventional studio portrait, a histogram shall show continuity of sampling and shall include values ranging from black to white. Some pixel values shall fall in the range of 5 - 12 (black) while some values shall fall in the range 243-250 (white).
- For color images with typical scene brightness that include white or black elements, similar values shall be provided for RGB (red, green, blue) renderings of white or black, i.e., a white area shall have values of r=243-250, b=243-250, and g=243-250 and a black area shall have values of r=5-12, b=5-12, and g=5-12.

C.4.4.2 Tonal Value (Desirable Archival Image)

The tonal value for the desirable archival images shall be as follows:

- For a black-and-white scene with a typical range of brightness, e.g., a landscape in daylight or a conventional studio portrait. If the desirable images are produced at 12 bits-per-pixel, the histogram shall show continuity of sampling and include values ranging from black to white, and some pixel values in the range of 5-15 (black) will be present as well as some values in the range 4070-4086 (white). If the desirable images are produced at greater than 12 bits-per-

pixel, the histogram shall show continuity of sampling and include values proportionally greater than those for 12 bits per pixel.

- For color images with typical scene brightness that include white or black elements, if the desirable images are produced at 36 bits-per-pixel, similar values shall be provided for RGB (red, green, blue) renderings of white or black, i.e., a white area shall have values of $r=4070-4086$, $b=4070-4086$, and $g=4070-4086$ and a black area shall have values of $r=0-10$, $b=0-10$, and $g=0-10$. If the desirable images are produced at greater than 36 bits-per-pixel, the histogram shall show continuity of sampling and include values proportionally greater than those for 36 bits per pixel.

C.4.5 Deriving the Reference and Thumbnail Images

The requirements for tonal representation shall apply to all three (archival, reference and thumbnail) images. The reference and thumbnail images, however, shall be reduced in scale and shall be sharpened, compressed, and color-indexed as applicable. The methods or techniques to be used to provide this additional image processing shall minimize image degradation and also produce derivative images that maintain the general look and character of the archival images.

C.4.6 Requirements Subject to Measurement

The image-capture system used to produce the images shall meet certain requirements that pertain to spatial resolution, tonal distribution, and noise (signal to noise ratio; meaning how many bits of the stored file are actual image information and how many bits are random noise of the system). Spatial resolution is determined by measuring the modulation transfer function (MTF) of the capture system. This will enable the Library to ensure that the delivered files have the required resolution and were not sampled up from a lower resolution; for example, an image produced from an 8x10-inch photograph at 3,000-pixels shall have a spatial resolution that is truly 300 dpi and was not sampled up from a 200 dpi file. The measurement of tonal distribution (C.4.6.2) shall confirm that the capture system is capable of capturing images that have the full range of tones as specified in the requirements above. The measurement of noise (C.4.6.3) shall ensure that the capture system produces a signal that includes as much image information as possible.

C.4.6.1 Modulation Transfer Function

The measured MTF shall have values which fall within the ranges given in the following table, at the given spatial frequencies:

Frequency	MTF
1	0.90 to 1.0
2	0.80 to 1.0
3	0.70 to 1.0
4	0.60 to 1.0
5	0.50 to 1.0
6	0.40 to 1.0
8	0.30 to 1.0

C.4.6.2 Tonal Resolution

For mandatory images (8-bit grayscale), the digital values should be linear to the original density. The digital values for each area on the grayscale target shall not deviate by more than 10 from a linear least squares regression line fitted between the densities of the original target and the digital output values. A white area shall have values of $r=243-250$, $g=243-250$, and $b=243-250$, and a black area shall have values of $r=5-12$, $g=5-12$, and $b=5-12$. Care should be taken that no clipping (= loss of details) in either the highlights or the shadows occurs.

For desirable images the digital values should be linear to reflectance / or transmittance. A white area shall have values of $r=4070-4086$, $g=4070-4086$, and $b=4070-4086$, and a black area shall have values of $r=5-15$, $g=5-15$, and $b=5-15$. Care should be taken that no clipping (= loss of details) in either the highlights or the shadows occurs.

C.4.6.3 Baseline Values for Noise, Flare, and Geometry

Measurement of the evaluation benchmark test (Section M.2.2) of the target images (Section C.4.6.4) will establish baseline values for noise, flare, and scanner geometry. These baseline values shall not deteriorate during the period of performance, i.e., noise and flare shall not increase nor geometry be adversely affected.

As indicated in Section C.7.5.2, the contractor shall include newly scanned images of the target set with each batch of images delivered to the Library, for use in determining that the baseline values are maintained. The Library will notify contractors if the values are not maintained. The contractor shall then take steps to bring the values to the levels measured earlier.

C.4.6.4 Targets for Objective Measurement

The conformance of the contractor's system to the preceding requirements shall be determined by measuring images of the target sets described below. The target sets and the measurement tools have been provided by the Image Permanence Institute (IPI) of Rochester, NY.

C.4.6.4.1 Transmitted light target set

The transmitted light target set will include the following elements:

A. Spatial resolution targets (produced by Sine Patterns Inc.).

- Sine Patterns M-6-60, size 46 x 70 mm (for larger format originals, used for Lot 1 evaluation)
- Sine Patterns M-7-60, size 21.5 x 30 mm (for smaller originals, used for Lot 2 evaluation)

The Sine Pattern target contains four rows of patterns. The top row contains seven different transmittance density patches. The next two rows contain sine wave patterns with different spatial frequencies, and the bottom row contains seven additional

different transmittance density patches. Each target is calibrated individually by the manufacturer.

B. Gray Scale targets. Measurement of the image of this target characterizes the relationship between the input values and the digital output values. In addition, the white, middle gray, and black area of the grayscale target will be used to measure the system noise (see C below).

- Target consisting of 12 different gray density patches on a middle gray background (for digital cameras).
- Target that includes a gray scale and a white, a gray, and a black area (for linear array scanners).

C. **An** Additional target.

- Knife-edge target for resolution measurement; this target will allow for an additional measurement of the spatial resolution using a different approach than the sine wave target. The target necessary consists of a slightly tilted sharp edge. The results can be compared to readings from the Sine Patterns targets (A. above); the knife edge target responds differently to image sharpening and improper resampling methods and helps determine if post-scan processing has been applied.
- **Additionally, the targets shall be evaluated for scanner noise, scanner geometry, and flare.**

C.4.6.4.2 Reflected light target set

The reflected light target set will include the following elements:

- A. Spatial resolution targets (produced by Sine Patterns Inc.).
- Sine Patterns M-13-60 (1x), size 47x70 mm.

The target contains four rows of patterns. The top row contains seven different reflectance density patches. The next two rows contain sine wave patterns with different spatial frequencies, and the bottom row contains seven additional different reflectance density patches. Each target is calibrated by the manufacturer.

B. Grayscale targets.

Same types as transmitted light target set above, except these are manufactured for reflected light readings.

C. **An** Additional target.

Same type as transmitted light target set above, except this has been manufactured for reflected light readings.

C.4.7 TIFF Header Requirements

TIFF version 5.0 shall be satisfactory; version 6.0 may be substituted as samples during project startup and is subject to acceptance by the Library. The Library uses the TIFF tags listed below. "Typical" or "expected" data are provided for most tags. Exceptions to the norm are noted in the comments column.

<u>Description</u>	<u>Tag</u>	<u>Comments</u>
NewSubfileType	254	
ImageWidth	256	actual pixel count
ImageLength	257	actual pixel count
BitsPerSample	258	
Compression	259	
PhotometricInterpretation	262	
DocumentName	269	collection identifier and filename* (Ex. bbc/0421ft.tif)
StripOffsets	273	
SamplesPerPixel	277	
RowsPerStrip	278	
StripByteCounts	279	
XResolution	282	**
YResolution	283	**
ResolutionUnit	296	**
DateTime	306	date and time scanned
Artist	315	Library of Congress

* Each collection identifier is a designation which shall also be used as part of the CD-ROM volume names (see C.9.2).

** At least two options exist for tags 282, 283, and 296, either one of which is acceptable:

Option 1 (often used for full-size uncompressed images)

Xresolution	282	actual pixel count
YResolution	283	actual pixel count
ResolutionUnit	296	1 (no unit specified)

Option 2 (often used for thumbnail images)

Xresolution	282	dots per inch
-------------	-----	---------------

YResolution	283	dots per inch
ResolutionUnit	296	2 (inch)

In order for the digital images to open in all software packages, the TIFF header tags shall be sorted into ascending numerical order.

C.4.8 Cropping

The Library wishes to provide researchers with a reproduction of the entire original item. Thus, images shall be framed and cropped to show the entire original item and beyond the item's edges. For negatives or other transmitted light items, each digital image shall reproduce that item's actual-image area, the border on the film that surrounds the image area, and a portion of the background (light box or scanner top) beyond the edge of the film. A similar approach shall be followed for reflected-light items; the whole print, whole mount, and a portion of the background (beyond the mount) shall be reproduced.

In the delivered images, the amount shown beyond the edge of the item shall be no less than 1.5 percent of the dimension of the long side image. Thus, for a 3,000 x 2,000-pixel image, the border beyond the reproduction of the original item shall consist of approximately 35 pixels on all four sides; for a 640 x 480-pixel image, the border shall consist of approximately 10 pixels on all four sides.

Exceptions to these requirements may be required for some collections as indicated in C.4.8.1 below and in Section J, Attachments 5 and 6 for LOT 2 collections.

C.4.8.1 Cropping - Multiple Items on a Single Mount

When multiple prints are mounted on a single board, e.g., eight Civil War portraits on one cardboard mount or three small prints mounted on a single scrapbook page, each image on a multiple-print mount shall be captured separately. Framing for each print shall extend beyond the image proper in a manner consistent with the cropping instructions in Section C.4.10. Specific task orders may additionally require the production of an image that captures the full physical item, i.e., a Civil War mount with eight portraits or a full scrapbook page with three prints.

C.4.9 Concatenation in Civil War Maps (LOT 2)

In order to achieve the required spatial resolution of the Civil War maps (or other historical maps) to be scanned from microfiche in LOT 2, the contractor may have to capture tiles or segments of the map images on the fiche and concatenate (stitch or join) these segments when producing the digital images to be delivered to the Library. See Section J, Attachment 6 (6.4.2) for specific requirements regarding concatenated images of the Civil War maps.

C.5 GENERAL NAMING REQUIREMENTS FOR FILENAMES, DIRECTORIES, AND ASSOCIATED DATABASES

In the Library's retrieval system, each catalog record includes a collection identifier (also called an "aggregate name") and a digital item identifier for the picture therein described. Together, the two identifiers are used by the retrieval system to fetch the image from the server in which the image is stored. The identifiers are considered the beginning and ending levels in a UNIX pathname, and the Library has established a set of "naming rules" to identify one or more directories that form the intervening (middle) levels of the pathname.

The delivered images will be loaded into the server by copying them en masse from the contractor's delivery disks. The contractor shall assign the correct name to each image and shall place sets of images into correctly named and organized directory structures in accordance with general specifications for determining the sets of identifiers and names outlined below and in accordance with detailed naming instructions to be provided for each collection.

C.5.1 Image Filenames

Excluding the images of the Civil War maps for which specific naming requirements are specified in Section J, Attachment 6, the filename assigned to each image, the first four to seven characters, shall consist of the *digital item identifier*. The final character before the "." shall indicate the image category (e.g., uncompressed archival version) and the filename extension shall indicate the image file format.

For each set of images (uncompressed archival image, compressed reference image, and thumbnail image) that reproduces the same source item, the filenames shall end as follows:

u.tif - for uncompressed archival files (types 5UA, 4UA, and 3UA; see C.4.3.1)

r.jpg - for JPEG compressed reference image files (type CRI; see C.4.3.2)

t.tif - for thumbnail image files (type THM; see C.4.3.3)

For the desirable images, the eighth character and extension shall be:

v.jpg = Very high resolution compressed display image (types 5DI, 4DI, and 3DI; see C.4.3.4)

The filenames for each image in a set shall begin with the same *digital item identifier*. For example, the digital item identifier for one of the negatives from the state of Alabama in the HABS/HAER collection is 414634p. The image files that reproduce this item shall be named 414634pu.tif (uncompressed archival image), 414634pv.jpg (desirable compressed display image, if offered), 414634pr.jpg (compressed reference image), and 414634pt.tif (thumbnail image).

C.5.2 Image Directory Structures

The *directory structure* which shall be created for the storage of the image files shall be as specified for each collection. For example, for the HABS/HAER collection, the

directory path, **al/al0800/al0897/photos**, shall contain the image files associated with the *digital item identifiers* 414631p through 414640p. An explanation of the elements in the pathname follows:

<i>al</i>	state of Alabama
<i>al0800</i>	range of lower-level directories (0800 through 0899)
<i>al0897</i>	HABS/HAER control number
<i>photos</i>	photo directory for Alabama 0897

C.5.3 Databases or Computerized Inventories

Naming instructions will specify whether a Library existing database shall be ungraded or a new database or computerized inventory is required. When naming instructions are provided for an existing database, new content shall be entered into the database. Section C.5.3.1 provides an illustrative example of such a database. Specific instructions will be provided for each collection.

When naming instructions require the contractor to examine marked items, the contractor shall create a new database enumerating the items and numbers assigned. In this database, the contractor shall enter the number or other mark on the original item in one field of a given data record and enter the file identifier used to name the set of image files

in another field. For some collections it may be required that these two data elements be linked to the identifier for the directory in which the images are stored.

For either an existing or newly created database, the contractor shall record notes that pertain to the production work at hand, e.g., a note about the condition or characteristics of the items being captured, or any other type of record keeping that may be useful in the production process. All information recorded in the database shall be included in the version delivered to the Library with the images.

Delivered databases shall be in a format capable of being loaded into common software, e.g., a comma-delimited ASCII file that can be loaded into Paradox or dBASE.

C.5.3.1 LOT 1- Collections With Pre-existing Databases

When the Library has a pre-existing database for the collection, it will be provided to the contractor in a format capable of being loaded into common software, e.g., a comma-delimited ASCII file that can be loaded into Paradox or dBase. Typically, the database will contain core information about the items or groups of items. The contractor shall add information to the database as images are captured, e.g., the names actually assigned to the files. As in the illustrative example below, some databases may contain single records that represent a group of item. In these cases, the contractor shall "clone" the record for each item actually captured before entering the identifier for that item (see notes for data elements 2, 3, and 7 in the example below).

For illustration, the following outline lists the fields in the HABS/HAER database and indicates how data shall be entered.

HABS/HAER Photograph Database

Field

Field Content

1. Control number (directory identifier), e.g., AL0897

In the database to be provided to the contractor. To be used by the contractor to name the directory for the images. The control number is unique in the database.

2. HABS survey number, e.g., AL-889

In the database to be provided to the contractor. Items to be scanned shall be marked with this number. See also number data element 7 below. The survey number is unique within this field of the database, but a survey number with the same value may appear as a HAER survey number also.

In this collection, there are about 25,000 HABS surveys; the database as delivered to the contractor will contain 25,000 HABS-related data records.

3. HAER survey number, e.g., MD-24

In the database to be provided to the contractor. Items to be scanned shall be marked with this number. The survey number is unique within this field of the database, but a survey number with the same value may appear as a HABS survey number also.

In this collection, there are about 10,000 HAER surveys; the database as delivered to the contractor will contain 10,000 HAER-related data records.

4. Number of negatives

In the database to be provided to the contractor. This is a net quantity for all sizes; no comparison of this number against the number of images actually captured can occur until all negative sizes have been scanned. This information may not be reliable and is being provided to the contractor for general guidance only.

5. Number of 4x5-inch color transparencies

In the database to be provided to the contractor. This information may not be reliable and is being provided to the contractor for general guidance only.

6. Library of Congress shelflist number.

All negatives and transparencies are sorted in shelflist order. The shelflist code will also appear on all negative sleeves and provides further item identification confirmation. The data will be sorted on this field to match the order of capture.

7. Digital item identifier for link to the digital image filenames, e.g., 414634p. This identifier shall be used to assign the names to the files 414634pu.tif, 414634pr.jpg, and 414634pt.tif.

Serial numbers shall assigned to each item by the contractor and added to the database. The identifier shall be established by and entered into the database by the contractor.

The contractor shall clone a new data record for each new item. The database will contain about 35,000 records (25,000 HABS and 10,000 HAER) when delivered to the contractor. The collection contains nearly 200,000 items and the final database returned to the Library shall contain nearly 200,000 individual data records, each with a unique number entered as the item-level identifier.

8. Item or photo number as written on negative or transparency storage sleeve, e.g., "2" from the full written entry "AL-889-2"

Shall be added to the database by the contractor at time of image capture. These numbers may repeat; the data shall be entered as written.

9. Note

Field for contractor to note line negatives or other anomalies as well as to copy any written information on negative sleeves that pertains to rights or restrictions.

10. Restrictions indicator field

To be checked when restriction information is

encountered.

11. Color indicator field

To be checked when color originals are scanned.

C.5.3.2 LOT 1- No Existing Database

The contractor shall refer to the *physical identification number* on each original item in order to determine the *digital item identifier*. The original items will be marked in some way with a unique number, e.g., files of negatives have numbers written on the envelopes or Civil War prints have numbers pencilled on the backs of the mounts. Instructions will be provided for the translation of the *physical identification number* into the *digital item identifier* which shall be completed at scan time. For example, for a copy negative in an envelope marked LC-USZ62-134356, the identifier will be 3c34356 and the three files to be created shall be named 3c34356u.tif, 3c34356r.jpg, and 3c34356t.tif. The *rule* for this example is that the last five digits of the physical identification number shall become the basis for the identifier, with a designated prefix to be added. Instructions will also be provided for naming directories to contain the files. For the above example, the contractor shall create the directory structure, 3c30000/3c34000/3c34300.

C.6 GENERAL HANDLING AND SCANNING REQUIREMENTS

C.6.1 LOT 1 - Photographic Negatives and Transparencies (transmitted-light items)

For LOT 1, the format for the bulk of the black-and-white photographic negatives to be scanned is medium-format (4x5 and 5x7 inches) safety film. The scan groups to be provided for the HABS/HAER collection will not be completely sorted by size, see Section J, Attachment 1.2.1. Other negatives range in size from 35mm to 8 x 10 inches. A portion of the HABS/HAER collection negatives have nitrate and/or diacetate film bases. Work with nitrate based film shall be completed in accordance with the special handling rules and requirements in Section D.3, page D-1.

Some of the 35mm strips of film are stored in PrintFile storage pages; and, in some cases, the entire PrintFile enclosure shall be scanned (plastic and negatives) to create a digital image that "looks like a contact sheet." Individual frames shall also be imaged, and these images shall be made through the plastic PrintFile material.

Color transparencies and color negatives range in size from mounted 2 x 2-inch slides to 8 x 10-inch sheet films. Color film materials are typically housed in mylar jackets or sleeves within an additional paper sleeve. Color slides are often housed in 20-slide translucent plastic racks. In the racks, the slides are held in place along their edges; and the image is fully exposed to view from the top. In some cases, it may be required that the entire plastic slide rack be imaged "like a contact sheet." To accomplish this, some ambient lighting shall be applied as a supplement to the lighting from below, in order to make the identification numbers written on the slide mounts legible.

All film-based materials, such as black-and-white photonegatives and color transparencies not in mylar sleeves shall be handled with cotton gloves and resleeved into their original housings. When rehousing, the emulsion side of film items shall face the non-sealed side (the side without an adhesive seam) of the sleeve or jacket.

C.6.2 LOT 1 - Photographic Prints and Other Graphic Print Materials (Reflected-light Items)

The established workflow shall be capable of accommodating item-to-item size variation.

Collections of prints and other printed graphic materials include items with great variation in size. For LOT 1, the smallest items to be captured include baseball cards approximately

1 x 2 inches and a pictorial button approximately 1-inch in diameter. No items larger than

11 x 17 inches shall be scanned.

Items in reflected-light collections are rarely grouped by size, i.e., there will often be considerable item-to-item size variation in a given collection. For example, in the Civil War prints, a 6 x 9-inch print on an 11 x 14-inch mount may be followed by eight 2.5 x 3.5-inch prints on a single 11 x 14-inch mount, followed by an unmounted 8 x 10-inch print in a mylar housing.

C.6.2.1 Condition of Reflected-light Items

Items identified as either fragile or having non-planar surface shall not (1) be flattened against or under glass or (2) turned face down for capture. The mounts are sometimes fragile and often non-planar, i.e., curved, cupped, or warped. The latter terms refer to deviations from a flat plane: Curved means that the mount is curved on one axis (the mount presents the appearance of a rocker in cross-section); cupped means the mount is curved on two axes; and warped means that the mount bears multiple or irregular distortions. Mounted stereographs, for example, are typically curved. The mounts may also be torn, broken, or brittle.

C.6.3 LOT 1 - Handling and Scanning

The capture device(s) to be utilized shall not cause harm to the materials being scanned. Harm may be caused by such factors as excessive handling, inversion of fragile items, flattening, surface abrasion, excessive illumination, and excessive heat.

The prohibitions on flattening and inverting non-planar reflected-light original items eliminate the universal use of flatbed scanners and glass plates to flatten items. Although flatbed scanners may be acceptable for the capture of transmitted-light items (and some reflected-light items), alternatives such as digital cameras that offer sufficient depth of field to keep non-planar items in focus shall be utilized for significant portions of this project. In addition, many items will be housed in mylar

sleeves and, in some case, shall not be removed for capture. Therefore, the scanning device shall be capable of capturing images through mylar.

C.6.4 LOT 2 - Original Materials

C.6.4.1 Civil War Map Fiche The Civil War map fiche consists of 105mm color Cibachrome microfiche. The capture device(s) shall not cause harm to the materials being scanned. Harm may be caused by such factors as excessive handling, surface abrasion, excessive illumination and excessive heat.

C.6.5 LOT 2 - Handling and Scanning

C.6.5.2 Civil War Maps on Fiche

The Civil War map fiche require care in shipping and handling. Shipment to and from the Library shall be made via an accepted overnight or next-day carrier that has been approved by the Library. The contractor shall be responsible for all shipping costs of pickup and return of Library materials and deliverables. The replacement cost for each fiche has been determined to be \$100.00.

C.6.5.3 Contractor Facilities

While on site at the contractor's facilities, the contractor shall store all Library property in a locked vault (except during periods of actual contractual work) and secure it from theft or damage.

C.7 WORKFLOW AND PROJECT MANAGEMENT

C.7.1 Contract Startup and Testing

Because of the complexity of the requirements and the variation in the Library's original materials, a startup and testing phase shall be the first task to be performed under contracts for both LOT 1 and LOT 2. The startup and testing phase shall provide a time during which the contractor and NDLP staff shall work together to address and finalize a mutually agreed upon definition of particular matters related to technical requirements and to confirm this understanding through sample work. Technical elements include but are not limited to:

- tonal resolution and color rendition
- cropping and orientation
- objective measurement of target images
- accuracy of filenaming and directory structure
- safety of the contractor's scanning system
- image sharpness from edge-to-edge for curved or warped materials

C.7.2 LOT 1 - Startup and Testing Phase

The HABS/HAER collection will be used as representative examples of the types of materials from which digital images shall be produced. The following materials accompanied by instructions regarding the filenaming and directory structure to be employed will be provided to the contractor:

1. 50 4x5-inch negatives from the HABS/HAER collection.
2. 20 8x10-inch negatives from the HABS/HAER collection.

C.7.2.1 Actions

The LOT 1 startup and testing phase shall include the following actions:

Initial Meeting	The COTR and other NDLP and Library staff members will meet with the contractor to discuss startup and testing activities and to present the 90 sample items.
System setup	The contractor shall set up and test the system(s) to be used for Lot 1. The time allotted for this setup shall be as proposed and agreed to prior to contract award.

Weeks after completion of system setup and testing:

Week 1	The contractor shall scan the 90 items, assigning names to files and placing them in directories as specified. The images shall be placed on a CD-ROM marked in accordance with delivery specifications. For the selections from the HABS/HAER, data shall be input into an existing database furnished by the Library.
Week 2	The images and associated database shall be delivered to the Library.
Week 3	The Library will complete quality review of the image samples and database and provide a brief preliminary written response concerning acceptability to the contractor.
Week 4	The contractor project manager and other contractor designated staff shall meet with the Library COTR and other NDLP staff to discuss the samples provided and to resolve any questions that may remain.

C.7.3 LOT 2 - Startup and Testing Phase

Two (2) rolls of original Mead/Bateson negatives and one (1) roll of original Mead/Bateson interpositives (also called diapositives) will be provided to the contractor.

C.7.3.1 Actions

The LOT 2 startup and testing phase shall include the following actions:

Activity start	The Library will ship three (3) rolls of original materials and a copy of the database to the contractor.
Week 1	after receipt The contractor shall examine the three rolls and meet with the COTR and other NDLP and Library staff members to discuss the various issues to be addressed during the startup and testing activities.
System setup	The contractor shall set up and test the system(s) to be used for LOT 2 in the time proposed and agreed upon prior to award.

Weeks after completion of contractor system setup and testing

Week	The contractor shall scan the 3 rolls, assigning names to files and placing them in directories in
------	--

- 1 accordance with the specifications. The items shall be placed on a CD-ROM disk that is marked in accordance with the delivery specifications. The contractor shall input digital item identifiers into the supplied database.
- Week 2 The contractor shall ship the images and the database to the Library.
- Week 3 The Library will complete a review of the image samples and database and provide a brief preliminary written response concerning acceptability to the contractor.
- Week 4 The contractor's project manager and other contractor staff meet with COTR and NDLP and other Library staff members to discuss the samples provided and to resolve any questions that may remain.

C.7.4 Task Orders

The production of digital images shall be performed under task orders issued under this contract. Task orders may include one or more jobs, i.e., separate and distinct digitization activities. For example, a single task order may include two jobs: (1) the digitization of a set of negatives from one part of the Library and (2) the digitization of a set of magazine covers from another part. Each job will be all or part of a specific *NDLP collection project* and may consist of multiple batches (see C.3).

C.7.4.1 Technical Preparation for Individual Task Order Jobs

Each job shall treat a coherent body of material and may require a technical setup phase prior to the issuance of a task order. During this phase, the COTR and the contractor's project manager shall resolve any outstanding issues or technical matters pertaining to the effort and shall establish other requirements, including but not limited to, the period of performance, delivery dates, batch sizes for deliverables, and the details of the database for the project. In addition, the contractor may be required to prepare a database format and provide 10-20 sample digital images that display technical solutions and levels of quality for the Library's inspection. Reimbursement for the sample images and technical labor for database preparation shall be in accordance with the contract, Section B, Schedule of Pricing.

C.7.5 Contractor Quality Control Program

A quality control program in accordance with the requirements for accuracy and delivery shall be initiated, documented, and maintained throughout the life of this contract. The Library expects that the contractor shall perform quality control for 100 percent of deliverables. A specific quality control plan shall be implemented for each phase of contract performance beginning with capture of images and ultimate acceptance by the Library of all deliverables. In addition, the contractor shall be responsible for inspecting the accuracy of filenames and directories for all digital images produced under this contract. Inspection hardware and software shall be of appropriate quality, accuracy, and quantity to ensure that all requirements of this contract are met.

The contractor shall document all quality control procedures, including actions taken to correct any problems, and submit a quality control report with or as a part of the

database with each delivery to the Library. This quality control report must enumerate and describe actions taken.

C.7.5.1 Contractor Quality Review: Imaging - LOT 1 and LOT 2

Contractor quality review shall include, but is not limited to, the following types of activities:

C.7.5.1.1 Image Quality

Acceptance criteria shall include but not be limited to a review of the following:

- complete item has been captured and proper cropping has been applied
- images are not skewed, blurred, indistinct, or flawed by dust or electronic artifacts or noise
- derivative images open and display properly
- images meet specifications for resolution, bit depth, level of compression, image orientation
- images of targets meet requirements for spatial and tonal resolution and do not show increase in noise

C.7.5.1.2 Other factors

Other factors related to the performance of the contract specifications shall include:

- capture and delivery of all items presented to contractor for scanning
- accuracy of filenaming, directory structure, CD-ROM volume naming
- completeness and accuracy of documentation in the form of tracking databases, file identification databases, and quality review reports

C.7.5.2 Objective Measurement - Test Targets

The Image Permanence Institute of Rochester, NY, has created test target sets for the Library of Congress; these are described in Section C.4.6.4. In order to permit the objective evaluation of the capture system as a part of the contractor's quality control program and by the Library's quality review, the contractor shall provide a scanned image of the target set appropriate to the job at hand with each delivered batch of materials.

C.7.5.3 Rework

Rework means the scanning of replacements for unacceptable digital images. For rework, the contractor shall follow all contract specifications and specific task specifications as agreed to for the original scanning and for the filename/directory structure, unless otherwise directed by the Library's COTR. (See Section C.9.6)

C.8 RELATED SERVICES

C.8.1 Programming and Processing Activities

The capability to provide different levels of technical expertise is required. It is anticipated that additional programming or processing steps associated with scanning or database requirements may be necessary. These tasks may require different levels of technical expertise which will be specified for task orders as applicable.

In addition to the preproduction analysis, these activities will fall into two general categories:

- Preparation of the *confirming samples* and other matters pertaining to the setup for each job listed in a task order.
- Actions that must be taken to manipulate images or other data because of exceptional factors or circumstances encountered as the work proceeds.

The contractor shall provide needed labor to carry out these related services by supplying a technician, computer programmer, or project leader as applicable.

C.9 DELIVERABLES AND DELIVERY

C.9.1 Types of Deliverables

The work to be performed shall yield deliverables of the types listed below. These will vary slightly from collection to collection.

The general types of deliverables shall include:

- Digital image sets delivered on suitable media. Each original item shall be reproduced by a set of three or four digital images (of varying specifications) and delivered on write-once CD-ROM disks. Each batch of images shall be accompanied by an image of the appropriate test target for the batch.
- Newly created or updated databases.
- Written documentation pertaining to the shipment to include a list of the delivered CD-ROMs with a printout of the directories and files.

C.9.2 Delivery Identification

A *disk name* shall be assigned to each CD-ROM used to deliver images. The disk name shall be composed of the *collection identifier* (to be provided by the Library) and a three-digit serial number starting with 001. This disk name shall be assigned as the computer-

encoded *volume name* for the disk (when the disk is formatted) and also written on the disk and its container with indelible ink. For example, disk/volume names for images in the Baseball Card Collection may be **bbc001**, **bbc002**, etc. and **habshaer001**, **habshaer002**, etc. for images in the HABS/HAER Collection

C.9.3 Delivery Media

Digital images shall be delivered on write-once CD-ROM disks compatible with all ISO 9660 specifications **except that the Library requires the use of lower-case letters in directory and file names**, in contravention of ISO 9660. Each CD-ROM and accompanying jewel case shall be labeled with the collection name, disk (volume) name, date completed, and the indicator *Library of Congress*.

C.9.4 Delivery Filenames and Directories

Each CD-ROM shall contain DOS files organized in DOS directories as indicated in the general guidelines in C.4 and the collection specific guidelines as provided.

C.9.5 Main Delivery/Alpha Disks

Delivery batches of one or more write-once CD-ROMs as the digital images are completed and written to CD-ROM, shall be shipped to the Library as *alpha* disks, defined as the first delivery of the image sets. The alpha disks will be retained by the Library.

C.9.6 Delivery of Rework

Unacceptable (rework) images shall be delivered on *rework disks*. If a rework batch consists of a small number of images, delivery may be on floppy disks or a new write-once CD-ROM. Separate floppy rework disks or rework CD-ROMS shall be produced for each collection (to facilitate archiving of the disks by the Library of Congress). Each rework disk shall be named and marked in a manner similar to that used for the main delivery disks, with the letter *r* added as the last character in the name and the word *rework* written on the disk label. Rework disks shall not contain any previously accepted image files.

C.9.7 Shipment Documentation

Each shipment of digital files on CD-ROMs shall be accompanied by directory and filename lists of the contents of each CD-ROM. The filename list shall contain file sizes and the date and time of creation information for each file.

C.9.8 Return of Government Furnished Materials

C.9.9 Replacement of Items

C.9.10 Intermediate Production Formats and Duplicate Digital Files

SECTION D

PACKAGING AND MARKING

D.1 PACKING AND MARKING

D.1.1 Delivery CD-ROM Disks

Requirements for marking CD-ROM disks and for the inclusion of accompanying documentation are provided in section C.9. When shipping CD-ROMs, the contractor shall adhere to the best commercial practices to resist breakage or other damage.

D.2 LOT 2 - RETURN OF LIBRARY ORIGINAL MATERIALS

Each shipping batch of LOT 2 materials shall be packed by the contractor for return to the Library to replicate the shipping batch sent to the contractor by the Library (see section C.3.3). The contractor shall pack shipping batches in separate labelled cartons that are sealed to provide protection against dirt, water, exposure to light, and physical damage in accordance with the best commercial practices which meets the packing requirements of the carrier and ensures safe delivery at the destination. Note that many of the Lot 2 materials are nitrate film; see section D.3 below.

If applicable, the cartons of deteriorated (delaminated) negatives shall be clearly labelled to distinguish them from the cartons of undeteriorated negatives or diapositives. See Section F.6 for the shipping address and proper procedures for delivery of materials.

The contractor shall provide the Library with an itemized packing list of negative numbers and a total for each carton of original negatives and corresponding cartons of black-and-white negatives, interpositives, duplicate negatives, and deteriorated negatives.

D.3 LOT 2 - SPECIAL HANDLING PROCEDURES - NITRATE FILM

The original film to be duplicated under the terms of this contract is cellulose nitrate, which is defined as a Class 4.1 Flammable Solid, a hazardous material. Specific Department of Transportation (DOT) and International Air Transport Association (IATA) laws and procedures include, but are not limited to, strict gross weight limits of film in cargo shipments; use of specific shipping containers for the film, notification of contents for carrier; and specific labelling requirements for shipments containing nitrate film. The procedures and regulations governing the shipment of cellulose nitrate film are subject to change without notice.

The contractor shall contact intended carriers to be used for shipments of the original film negatives, and shall provide to the Library a written procedure demonstrating both a willingness to comply and general understanding of applicable regulations as they pertain to the conveyance of nitrate flat film. The contractor must procure the proper shipping containers, labels, and shipping forms for purposes of returning original nitrate negatives to the Library.

SECTION E
INSPECTION AND ACCEPTANCE

E.1 NOTICE LISTING CONTRACT CLAUSES INCORPORATED BY REFERENCE

NOTICE: The following solicitation provisions and/or contract clauses pertinent to this Section are hereby incorporated by reference:

FEDERAL ACQUISITION REGULATION (48 CFR CHAPTER 1)		
52.246-02	INSPECTION OF SUPPLIES - FIXED PRICE	AUG 1996
52.246-04	INSPECTION OF SERVICES - FIXED PRICE	AUG 1996

E.2 INSPECTION AND ACCEPTANCE

The Library of Congress reserves the right to have the Contracting Officer and/or designated COTR inspect the contractor's facilities during the actual production of digital files, including work and storage areas, whether these areas be located at the contractor site or on Library premises.

The contractor is responsible for performing all inspections or evaluations of the quality of all digital files and the correctness of digital file and directory names during production and prior to delivery to the Library.

All unacceptable individual images shall be corrected at no additional cost to the Library. Unacceptable individual images may be identified at the time of delivery and initial inspection or at any time during the period of performance. All unacceptable batches shall be corrected at no additional cost to the Library. Unacceptable batches will be identified at the time of delivery and initial inspection, following the procedure outlined below.

The Library of Congress will require two (2) weeks to perform inspections and to conduct tests to determine acceptance for each delivered batch. In addition, the Library will require eight weeks at the end of each completed task order to carry out a final wrap-up review of the final collection and the overall project.

E.2.1 Acceptance Procedures

E.2.1.1 Image Acceptance Criteria

The images shall meet the general requirements outlined in Section C and the specific requirements outlined in the attachments to Section J and in separate task orders. Acceptance criteria for images shall include but not be limited to the following:

- complete item has been captured and proper cropping has been applied
- images meet specifications for tonal range
- images open and display properly

- images meet specifications for format, resolution, bit depth, level of compression, image orientation, and header content
- images are not skewed, blurred, indistinct, or flawed by dust or electronic artifacts or noise

E.2.1.1.1 Inspection of Mandatory Images (8-bit and 24-bit)

When evaluating the mandatory images (8 bits-per-pixel for grayscale, 24-bits for color), the Library will view them on a display monitor that has been calibrated using the National Archives and Records Administration monitor calibration target. (Copies of this target will be provided to offerors upon receipt of written requests submitted to the Contracting Officer.) The Library will view the images and inspect the histograms to ensure that the requirements regarding tonal values (C.4.4.1) have been met.

E.2.1.1.2 Inspection of Desirable Images (minimum 12-bit and 36-bit)

When evaluating the desirable images (minimum 12 bits-per-pixel for grayscale, 36-bits for color), the Library will examine histograms for the images to ensure that the requirements regarding tonal values (C.4.4.2) have been met. The Library will also view the display images (5DI, 4DI, and 3DI) that were derived from the archival image on a display monitor that has been calibrated using the National Archives and Records Administration monitor calibration target. (Copies of this target will be provided to offerors upon receipt of written requests submitted to the Contracting Officer.) The Library will examine the desirable display images and their histograms to ensure that the requirements regarding tonal values (C.4.4.1) have been met.

E.2.1.2 Batch Acceptance Criteria

Delivery batches shall meet the general requirements outlined in Section C and the specific requirements outlined in the attachments to Section J and in separate task orders. Acceptance criteria for batches shall include but not be limited to the following:

- all items presented to contractor for scanning have been captured and delivered
- no more than the permitted number of images in the sample set (see E.2.2 below) fail to meet the image acceptance criteria
- target image supplied with batch indicates capture system meets requirements
- directory names and filenames assigned correctly
- database created or enhanced correctly
- delivery media has correct volume names and properly marked packaging

E.2.2 Sampling and Acceptable Quality Level for Batches

The Library will select sample images from each batch in accordance with the procedures outlined in the American National Standard system (ANSI/ASQC Z1.4-

1993 and ANSI/ASQC S2-1995). The Library will use General Inspection Level II and Acceptable

Quality Level (AQL .65), i.e., the size of the sample will be determined by the size of the batch and calculated according to the sampling plan for Level II, AQL .65.

As quality review proceeds during the life of this contract, the Library will use the switching rules as stated in the standard. For example, normal inspection will be followed for the first ten (10) inspection batches. If no batches are rejected, reduced inspection rules will then be followed; if a batch is rejected, tightened rules will be followed. The following table illustrates the circumstances under which a hypothetical batch would be rejected:

E.2.2.1 Normal Inspection

Inspection Batch	Selected sample	Reject Batch
2,500 images	125 images	3 failed images

E.2.2.2 Tightened Inspection

Inspection batch	Selected sample	Reject batch
2,500 images	125 images	2 failed images

E.2.2.3 Reduced Inspection

Inspection batch	Selected sample	Reject batch
2,500 images	50 images	3 failed images

**SECTION F
DELIVERIES OR PERFORMANCE**

F.1 NOTICE LISTING CONTRACT CLAUSES INCORPORATED BY REFERENCE

F.2 PERIOD OF PERFORMANCE

F.3 DELIVERABLES

F.4 TIME OF DELIVERY

F.5 SCHEDULING DELIVERIES

F.6 PLACE OF DELIVERY (F.O.B. DESTINATION)

**SECTION G
CONTRACT ADMINISTRATION DATA**

G.1 INVOICES

G.2 PAYMENT DUE DATE

G.3 TECHNICAL DIRECTION

G.4 PAYMENT METHODS

SECTION H SPECIAL CONTRACT REQUIREMENTS

H.1 RELEASE, PUBLICATION, AND USE OF GOVERNMENT FURNISHED DATA

H.2 INTERPRETATION OF CONTRACT REQUIREMENTS

H.3 CONTRACTOR COMMITMENTS, WARRANTIES, REPRESENTATIONS

H.4 USE OF LIBRARY OF ENGINEER NAME OR CONTRACTUAL RELATIONSHIPS IN ADVERTISING

H.5 NEWS RELEASE

H.6 CONTRACTING OFFICER'S TECHNICAL REPRESENTATIVE (COTR)

H.7 KEY PERSONNEL REQUIREMENTS

H.8 REPRESENTATIONS AND CERTIFICATIONS

SECTION I CONTRACT CLAUSES

SECTION J - ATTACHMENT 1

HISTORIC AMERICAN BUILDINGS SURVEY

HISTORIC AMERICAN ENGINEERING RECORD REQUIREMENTS

1.1 BACKGROUND

This project will create digital image reproductions of the visual documentation compiled by the National Park Service for the Historic American Buildings Survey (HABS) and the Historic American Engineering Record (HAER). The goal of the collections is to provide architects, engineers, scholars, and the public with comprehensive documentation of buildings, sites, structures and objects significant in American history and the growth and development of the built environment.

1.2 GOVERNMENT FURNISHED MATERIALS

Each black-and-white photographic negative and color transparency is housed in a paper sleeve, marked with a unique *item identification number*. The identification number shall be used to find the associated database record.

1.2.1 Black-and-white Photographic Negatives

The following represents an estimate of the black-and-white photographic negatives in the collection as of January 1997. This represents work from 1933 forward; additions are made to the collection each year.

<u>Size</u>	<u>Quantity</u>
8 x 10	8,180
5 x 7*	50,190
4 x 5*	95,500
2 x 2*	2,340
Other**	4,750

* Stored and interfiled in 5 x 7 sleeves.

** Most other-size negatives range from 2 x 3 inches to 2 x 4 inches; the category also includes a few negatives as small as 1-inch square and others with high aspect ratios (2x6-inch).

Scanning shall be done in batches of negatives; each batch will represent a block of negatives that are stored in same-size **sleeves**. At least ninety percent of the negatives are stored in 5 x 7-inch sleeves that actually contain 5x7-inch, 4x5-inch, 2-inch, and other non-standard film sizes. Thus many batches of negatives to be scanned will not consist of same-size **negatives**.

The negatives are sorted in the Library's shelflist number order, which is geographically based. There is considerable variation in the age, physical condition, type of film, processing methods, and the size of negatives that are stored together; therefore, scanning adjustments shall be required while proceeding from one set of negatives to the next. NOTE: Approximately 18,200 of the negatives have a nitrate film base. See section 1.5 below.

The 8x10-inch series includes some lithographic line negatives. The system used for continuous tone, full-range negatives may not perfectly reproduce these essentially bitonal items. The best means conveniently capable and available shall be used to capture the line negatives and the presence of each shall be indicated in the "scanning note" field of the database (see section 1.4.1 below)

The HABS and HAER negatives are filed separately from one another and shall be captured at different times.

1.2.2 Color Transparencies

The following represents an estimate the variety of color transparencies in the collection:

5 x 7	280
4 x 5	1,250

In addition to the paper sleeve, the transparencies are stored in a mylar sleeve or jacket. There is less variability in the image quality of the color transparencies because HABS/HAER did not begin using color as a film format until the late 1970s.

1.3 IMAGING REQUIREMENTS

In accordance with C.4.3, the following set of mandatory images shall be created for each item in the collection:

5UA - 5000-Pixel Resolution Uncompressed Archival Image

CRI - Compressed Reference Image

THM - Thumbnail Image

If applicable, the desirable image type to be created shall be as follows:

5DI - 5000-Pixel Compressed Display Image

1.4 SPECIFIC REQUIREMENTS FOR FILENAMES, DIRECTORIES AND THE HABS/HAER DATABASE

1.4.2 File Naming

The contractor shall assign each digital image file a unique file name composed of the *digital item identifier* followed by the image category indicator and the file extension (see section C.4.2).

1.4.3 Directory Structure

1.4.3.1 Elements of the Directory Structure

1.4.3.2 Path Examples

1.5 HANDLING AND SCANNING REQUIREMENTS

The negatives and transparencies in the collection shall be handled with white cotton gloves. Although, the Library has not discovered any negatives or transparencies exhibiting signs of deterioration, any such items discovered by the contractor shall be brought to the attention of Library staff.

The approximately 20,000 nitrate and diacetate based films are currently stored in individual sleeves and are housed in a cold storage environment. These materials will be brought to the contractor in large grey coolers with clasp locks. The scan groups for these materials will be small enough so that the negatives may be stored in sealed lockers at night. For additional information on handling nitrate and diacetate materials, see Section J, Attachment 5.

1.6 WORKFLOW AND SETUP ACTIVITY

The workflow and setup for this collection are included in the contract startup and testing activity; see Section C. 6.2.

1.7 DELIVERABLES AND DELIVERY

The *collection identifier* for the HABS/HAER Collection is **habshaer**. This *collection identifier* shall be used as the first part of the volume name for the delivered CD-ROMs. See Section C.8.2

SECTION J - ATTACHMENT 2

CIVIL WAR MAP FICHE GROUP - LOT 2

2.1 BACKGROUND

The project described in this section is being carried out by the Geography and Map Division and the National Digital Library Program. The purpose of this project is to digitize approximately 2,000 Civil War maps that have been placed on approximately 3,000 105mm color microfiche. Many of the larger maps have been divided into segments or sections; each segment or section was separately photographed for the fiche. This segmentation explains why the quantity of fiche exceeds the number of maps.

The maps on fiche represent about one-fifth of the Library's total holdings of Civil War maps. The entire corpus, including the maps reproduced on fiche, is described in Richard W. Stephenson's book *Civil War Maps: an Annotated List of Maps and Atlases in the Library of Congress* (Second Edition. Washington: Library of Congress, 1989. ISBN 0-8444-0598-1, for sale by the Superintendent of Documents, Government Printing Office). This volume describes the original maps and lists the *bibliographic entry number* (i.e., catalog number) assigned to each map.

The original maps range widely in size: the smaller maps include examples at 10x7 cm (ca. 4x3 inches), larger maps include examples at 116x84 cm (ca. 45x33 inches). Most of the larger maps have been filmed in segments. The precise range of segment sizes is not known; a partial survey suggested that few fiche images represent map segments greater than 100 cm (ca. 40 inches); many appear to be in the range of 40-70 cm (ca. 15-30 inches).

The Library wishes to create digital archival or "master" images of these maps and map segments. Within the limits of fiche-scanning technology, these archival images are intended to match the quality and scale of the archival files that the Library produced when it scans similar original paper maps directly. These files are generally 300 dpi, 24-bit color uncompressed images at the actual scale or size of the original. For example, a map segment measuring 35x44 cm (ca. 14x17 inches) would be represented in a digital image with a resolution of about 4200x5100 pixels.

After receiving the images from the contractor, Library will archive them in its digital repository and create copies in one or more formats that will provide researchers with online access to the collection.

2.2 GOVERNMENT FURNISHED MATERIALS

2.2.1 Film and Condition

The contractor shall create digital images of approximately 3,000 maps or map segments from the Civil War maps on fiche group. Each map or map segment is the sole content of a 105mm fiche. A 105mm fiche measures 4 3/8 x 6 inches. The images of the maps or map segments on the fiche are generally about the same vertical size (roughly 3 to 3.5 inches), with the horizontal dimension falling where it may, but not exceeding about 5 inches.

In addition to the map or map segment, the fiche image includes (1) a six inch ruler and (2) a set of Kodak color control patches. In order to have the map fill the frame (to the degree possible), the fiche were made at varying reduction ratios.

The fiche are on Cibachrome stock and represent the camera original, i.e., the actual stock exposed in the fiche camera. The film was exposed by Library staff and processed by MicroColor International of Midland Park, New Jersey.

2.2.2 Item Identification

Each fiche is marked across the top with an identification that includes:

- the map title or its equivalent
- the bibliographic entry (catalog) number of the map
- indication of segmentation, e.g., *1 of 2, 2 of 3*
- indication of copy number when the Library holds more than one copy, e.g. *c.2* for copy 2

The information on the fiche will be the basis for assigning the filename to the digital image of the map or map segment.

2.3 ADDITIONAL IMAGING REQUIREMENTS

2.3.1 Determining Image Pixel Dimensions

The pixel dimensions of the delivered images shall be the same as the pixel dimensions would have been if the original paper item had been scanned at 300 dpi. The contractor shall determine the proper pixel dimensions by analyzing the ruler pictured on the fiche. After the captured image has been properly scaled, the length of the six inch ruler shall be represented by 1800 pixels (6 inches x 300 dpi = 1800). The pixel dimensions of the delivered images shall not vary from the values calculated for the ruler by more than 5 percent. For example, for a map with a theoretical vertical

dimension of 4200 pixels, the actual vertical dimension of the delivered image must fall between the values 3927 (4134 less 5 percent) and 4341 (4134 plus 5 percent). Allowing for this variance, the length of the six-inch ruler shall be represented by not less than 1710 nor more than 1890 pixels.

The pixel dimensions of the delivered images shall represent the actual optical resolution of the capture device.

2.3.2 Concatenation of Images

The original maps and map segments range to the order of 30 inches and thus must be reproduced in digital files that range to the order of 9000 pixels. If the contractor's capture device offers resolution lower than that required, then the contractor shall concatenate a composite image from a series of scans of tiles or segments of the fiche image.

The concatenated images shall minimize the evidence of the edges where tiles or segments are stitched or joined. The most important factor is the alignment and scaling of the segments; the Library requires that lines or words on the parts of the map image be in alignment. The second factor is matching of shade or tone; the Library requires that RGB values for a single uniform area of the original map that fall into more than one tile or segment (before stitching) be within 5 percent of each other.

2.3.3 Header Content

The general requirements for TIFF headers given in Section C.4.3 apply to these images. For clarification, the Library states that the headers for these files shall represent the values for width, length, and resolution that would have been obtained if the entire map or map segment had been captured from paper as a single image. For example, in the case of the first map in the table above, the width would be 5198, the length 4134, the resolution 300, and the resolution unit dpi.

2.3.4 Cropping and Borders

The contractor shall include in the delivered image:

- the entire map or map segment as shown on the fiche
- the ruler, generally photographed lying to the left of the map
- the color control patches and gray scale (if any), generally photographed lying to the right of the map

These items shall be captured in a *capture rectangle* that goes beyond the edge of the map, ruler, or color control patches. The capture rectangle shall always be beyond or outside the map, ruler, or color control patches. In effect, the capture rectangle forms a *border* around the objects; this border shall never be less 20 pixels in width. The color of the capture rectangle/border shall be whatever color is present on the microfiche; that is, there is no requirement to make the rectangle/border precisely the same shade for each digital image.

2.4 FILENAMING AND DIRECTORY STRUCTURE

The contractor shall assign filenames to the map images based upon the information printed on the top of the fiche. The files shall be in the TIFF format. Since for a portion of their life cycle at the Library, the images will be processed in DOS system computers, the filenames shall be eight characters long, with the extension *.tif*. Since so few map images will fit on each CD-ROM, there are no requirements for creating named directories.

The assignment of names by the contractor shall adhere to the following guidelines:

Char	Assigned content	Example	Fiche mark at top
1	Bibliographic entry number before the separator point ("decimal point"); if needed, fill before and after with zeros to make five character expression	05330	533
2			
h0810		H81	
3			
0520a		520a	
4			
s0420	S42		
5			
0066b	66b.5		
6	Bibliographic entry number after the separator point ("decimal point"); if needed, fill after with zeros to make two character expression	.20	465.2
7			
.22		456.22	
8	Feature code (see below)	d	2 of 3

Feature codes. The eighth character shall consist of a code to represent a feature of the map being scanned. The codes are as follows:

- a [Segment] 1 of 2
- b [Segment] 2 of 2
- c [Segment] 1 of 3
- d [Segment] 2 of 3
- e [Segment] 3 of 3
- f [Segment] 1 of 4
- g [Segment] 2 of 4
- h [Segment] 3 of 4
- i [Segment] 4 of 4
- x Copy 1
- y Copy 2
- z Copy 3
- 0 no feature

Examples:

Filename Mark on fiche

0066b50b 66b.5 2 of
0273000y 2
h268000a 274 c. 2
s1500000 H268 1 of
2
S150

2.5 DELIVERABLES

In accordance with C.4.3, the following set of images shall be created for each individual fiche in the collection:

CUA - Cartographic Uncompressed Archival Images

2.6 WORKFLOW AND SETUP ACTIVITY

The workflow and setup for this collection will entail scanning and delivery of ten samples from the collection.

The scanning, follow-on rework, and return shipments of negatives shall be completed within about 40 weeks. Each batch will consist of about 300 fiche; the entire collection includes 10 batches. At no time will more than 900 fiche be away from the Library; one batch en route, one batch at the contractor, and one batch being returned. Additional information on the movement of materials between the Library and the contractor found in Section F.

2.7 DELIVERABLES AND DELIVERY

The *collection identifier* for the Civil War fiche group is **cwfgm**. This *collection identifier* shall be used as the first part of the volume name for the delivered CD-ROMs. See Section C.9.2

PART IV - REPRESENTATION AND INSTRUCTIONS

SECTION K REPRESENTATIONS, CERTIFICATIONS AND OTHER STATEMENTS OF OFFERORS

SECTION L

INSTRUCTIONS, CONDITIONS, AND NOTICES TO OFFERORS

L.1 FORMAT AND INSTRUCTIONS FOR PROPOSAL

L.2 VOLUME I - STANDARD FORM OF CONTRACT AND PRICE PROPOSAL

L.3 VOLUME II--TECHNICAL/MANAGERIAL PROPOSAL

SEPARATE PROPOSALS (INCLUDING BOTH COST AND TECHNICAL) MUST BE SUBMITTED FOR EACH LOT PROPOSED. Offerors shall clearly state that the proposal submitted is in response to LOT 1 or LOT 2. The requested responses to L.3.1 through L.3.7 are required for both LOT 1 and LOT 2 unless specifically noted.

Organization of responses in Volume II shall be submitted in the order listed below. Comprehensive responses to the requirements of the Request for Proposals are necessary to evaluate the offeror's capability to meet the stated requirements and provide the deliverables described in the solicitation. Technical proposals should be practical, legible, clear, and coherent. In order that evaluation may be accomplished strictly on the merit of the material submitted, no costs shall be included in technical proposals.

General statements that the offeror can comply with the requirements will not, by themselves, be adequate. Failure to provide the requested technical information in L.3.1 - L.3.7 that follow, may be cause for rejection of the offer.

L.3.1 SECTION 1 - Overall technical approach; proposed methodology; demonstrated understanding of the scope of work and the requirements

The offeror shall address each of the requirements as listed in Section C. Detailed responses to each of these requirements will provide an explanation indicating offeror's ability and proposed methodology to be utilized to meet each requirement. Responses are not be a restatement of the requirement but shall be comprehensive, well-conceived, and include detailed approaches to accomplishing the tasks and providing the deliverables.

The offeror shall include specific responses which demonstrate the capability and proposed methodology to meeting the requirements, as listed in the following sections.

L.3.1.1 Mandatory Image Requirements

L.3.1.1.1 Image Capture System

LOTS 1 and 2

- Describe in detail each hardware and software system to be employed, including overall flow of images from capture and naming, to post-processing, quality review, and production of delivery CD-ROMS.
- Describe types of lighting to be used, appropriateness to the capture devices proposed, and characterize the lights in terms of the amount of heat produced.

LOT 1

- Discuss how the system(s) shall accommodate items that may be curved, cupped, or warped.
- Discuss how multiple capture stations shall be operated at the same time.

LOT 2

- Discuss how the system shall handle the three-foot-long rolls of 35mm film.

L.3.1.1.2 Technical Management

LOTS 1 and 2

- Describe how workflow shall be managed to meet schedules and provide a workflow chart in accordance with delivery requirements.
- Describe facility resources for the provision of related services including programming and custom processing.
- Describe approach to be used to perform the startup and test activity. Provide a calendar indicating how long equipment installation and setup will take prior to beginning the startup and test activity.
- Discuss the requirements pertaining to the handling of Library materials to indicate understanding of the problems and issues involved.
- Explain how the cropping requirements shall be met.
- Discuss methods for overall production control, including method for tracking the individual physical items provided by the Library and the digital images produced.
- Discuss format and other requirements for delivery media to indicate understanding; describe methods to be employed to name and mark delivery media.

LOT 1

- Describe how the capture activity shall be configured to fit the space provided.

LOT 2

Discuss the proposed plan for pickup and delivery of the original film; provide specific details concerning intended shipper(s), pickup procedures, etc.

L.3.1.1.3 Images and Imaging

LOTS 1 and 2

- Describe the method to be used to produce the various mandatory image types, including discussion to indicate understanding of digital image resolution, file format and file headers, and compression requirements
- Discuss the imaging implications caused by variation in the characteristics of the print, negative, and transparency types to be scanned.
- Describe the method for creating the TIFF header, including the content for the three special tags that identify the Library and provide the date and image identifier.
- Discuss the issues pertaining to the tonal resolution required for the mandatory images, indicating understanding of the issues and the requirement, and describing the technical means to be used to meet the requirement and judge the results.

- Discuss the distribution of tones and colors for the mandatory digital images, including a discussion of output devices, especially display monitors. How does use of a display monitor in viewing and judging the quality of these images influence decision-making about the distribution of tones or colors in a digital image? Discuss the advisability of "contrast stretching" or other adjustments to tonality, and the use of histograms and other tools in the judgement of image quality.
- Describe the methods to be used to calibrate the contractor's display monitor(s) at the capture station and at quality review workstations.
- Discuss the methods to be used to produce the derivative CRI (compressed reference) and THM (thumbnail) images, including how they will be compressed, rescaled, sharpened, and color-indexed (8-bit color thumbnails).
- Discuss the objective-measurement target, indicating an understanding of this requirement for reflected- and transmitted-light materials, and explaining how the required images will be produced.

LOT 2

- Discuss how the images of Civil War maps from fiche will be concatenated

L.3.1.1.4 Filenaming and Associated Databases

LOTS 1 and 2

- Provide discussion which indicated understanding of filenaming and database requirements.
- Describe the methods, staff, and tools to be used for assigning the required names for files and directories and for entering required data elements in the associated databases. Explanation of whether the assignment of names will be part of the initial capture process, post-processing, or both.

L.3.2 SECTION 2 - Project Management and Qualifications/Experience of Key Personnel

L.3.3 SECTION 3 - Previously Demonstrated Experience and Successful Past Performance

L.3.4 SECTION 4 - Quality Control

The offeror shall submit a detailed Quality Assurance Plan in accordance with and reflective of an understanding of the ANSI/ASQC Z1.4-1993 Standards. In the plan, the offeror shall describe the procedures and methods and staffing to be used to review the digital images and file/directory names before delivery to the Library to ensure that the delivery requirements are met. The plan shall address quality control procedures for handling reworks for unacceptable batches and individual images.

L.3.5 SECTION 5 - Corporate Support Capabilities and Facilities

LOT 2

Offeror shall describe the features of the contractor facility pertaining to the handling and storage of nitrate film which demonstrates the capability to maintain and assure the safety of the original materials.

L.4 SAMPLE IMAGES AND BENCHMARK TESTS

After the initial technical evaluation, and as indicated in Section M.2.2, the Library will require a demonstration of technical competence from those offerors determined to be in the competitive range (technical, price, and other factors considered). This technical demonstration will consist of the production and delivery of digital images. For both LOT 1 and LOT 2, the offeror will be required to produce a series of images.

For LOT 1, the Library will supply two special targets (for transmitted and reflected light) and **four** pictorial source images as described in Section M.2.2.2.

For LOT 2, the Library will supply one special target (for transmitted light), **two** five-frame strips of 35mm film and two Civil War map microfiche as described in Section M.2.2.2.

The offeror shall create images of the targets for evaluation by an independent testing laboratory and image sets as required for the **four** pictorial item for evaluation by the Library. The sample digital images together with any explanations that the offeror feels would be appropriate to understand the results shall be submitted for evaluation. **Two copies of the** digital images shall be submitted on write-once CD-ROM disks.

SECTION M

EVALUATION FACTOR FOR AWARD

M.1 EVALUATION CRITERIA

M.1.1 Contractor selection will be based on evaluation of proposals in accordance with the responses received to the criteria outlined in Section L, Instructions, Conditions, and Notices to Offerors and the Schedule of Prices. Award will be made to that offeror whose combination of technical and price proposals represents the best value to the Government and is most advantageous, price and other factors considered, and which is within the available Library of Congress resources.

M.1.2 The Library of Congress also reserves the right to reject any or all proposals received and/or request clarification or modification of proposals. The Library reserves the right to determine a competitive range for negotiation based upon the technical and cost acceptability of proposals. In addition, the Library reserves the right to award a contract without discussions.

M.1.3 Cost evaluation will include an analysis of the total cost and cost elements (if applicable) to perform the required work. The total costs supplied by the offeror shall be submitted on a copy of Section B in the spaces provided and shall constitute the total firm-fixed unit price for that service or deliverable.

M.1.4 Proposals that are unrealistic in terms of technical commitment or unreasonably low or high in cost or price will be deemed reflective of an inherent lack of technical

competence or indicative of failure to comprehend the complexity and risk involved in the contract requirements and may be grounds for rejection of the proposal.

M.2 EVALUATION FACTORS

Technical proposals will be initially evaluated with respect to four (4) major factors for determination of the competitive range. Technical factors are listed in descending order of importance. The technical proposal is worth more than the cost proposal; when technical proposals are relatively equal in technical merit, cost may increase in importance.

M.2.1 Technical Factors

- Factor 1** Overall technical approach; proposed methodology; demonstrated understanding of the scope of work and the requirements (L.3.2)
- Factor 2** Previous demonstrated production experience and past performance; qualifications of key personnel and project management (L.3.3)
- Factor 3** Quality Control (L.3.4)
- Factor 4** Facilities and corporate support capability (L.3.5)

M.2.2 Sample Digital Images and Benchmark Tests

Those offerors determined to be in the competitive range (technical, price, and other factors considered) shall be required to provide images of special technical targets and of pictorial items as indicated below. The sample images for both Lot 1 and Lot 2 will be evaluated on a pass/fail basis in terms of the considerations outlined in Sections M.2.2.1 (targets) and M.2.2.2 (pictorial images).

M.2.2.1 Images of Target Sets

For LOT 1, offerors in the competitive range shall produce images for two (2) technical target sets (one transmitted-light target set and one reflected-light target set). The target sets and complete instructions regarding procedure will be provided to the offerors for the benchmark test. The target sets are described in Section C.4.6.4; each set includes: (a) a spatial resolution target, (b) a grayscale target, and (c) an additional target to characterize the scanning system. For LOT 2, images of only the transmitted-light target set will be required. The images will be evaluated by an independent testing laboratory.

M.2.2.1.1 Pass-fail values for spatial resolution

The readings from the Sine Patterns sinusoidal target shall be the same for reflected and transmitted light. The measured MTF shall have values which fall within the ranges given in the following table, at the given spatial frequencies:

Frequency	MTF
1	0.90 to 1.0

2	0.80 to 1.0
3	0.70 to 1.0
4	0.60 to 1.0
5	0.50 to 1.0
6	0.40 to 1.0
8	0.30 to 1.0
10	0.20 to 1.0

M.2.2.1.2 Pass-fail values for tonal distribution

The measurements from the grayscale targets shall be the same for reflected and transmitted light.

For mandatory images (8-bit grayscale), the digital values should be linear to the density of the original. The digital values for each area on the grayscale target shall not deviate by more than 10 from a linear least squares regression line fitted between the densities of the original target and the digital output values. A white area shall have values of r=243-250, g=243-250, and b=243-250, and a black area shall have values of r=5-12, g=5-12, and b=5-12. Care should be taken that no clipping (= loss of details) in either the highlights or the shadows occurs.

REQUIREMENT FOR IMAGE SAMPLE FOR THE DESIRABLE DIGITAL IMAGES OMITTED FROM BENCHMARK TESTING

M.2.2.1.3 Crosscheck evaluation of additional targets and sample pictorial images

In order to confirm the findings in the preceding tests and in order to offer diagnostic description of the offeror's capture system, the testing laboratory will make some additional measurements. Some additional measurements will be made from targets scanned for spatial resolution or tonal distribution; some will be made from additional targets; and some will be made by examining the pictorial samples provided to the Library evaluation committee (Section 2.2.2). The crosscheck measurements will include the following:

- White, middle gray, and black area (same for reflected and transmitted light). Measurements of these zones on the Gray Scale will indicate system noise (signal-to-noise measurement).
- Flare from the flare measurement target.
- Confirmation of spatial resolution measurement (from the Sine Patterns target) by measuring the results from a knife-edge target.
- Measurement of scanner noise from 1) a measurement without a sample in the light path and 2) measuring the dark current (i.e., making a scan with the lens covered in the case of a camera; or scanning a black target if the light source cannot be turned off).

- Measuring of scanner geometry related source noise amplification.

M.2.2.2 Images of Pictorial Items

LOT 1

Sets of digital images for **four (4)** pictorial items as follows shall be produced--

1. Black-and-white 8x10-inch negative. Produce 5UA, CRI, THM mandatory image set and 5UA, 5DI, CRI, THM desirable set, if offered.
2. **DELETED**
3. Color 4x5-inch transparency. Produce 4UA, CRI, THM mandatory image set and 4UA, 4DI, CRI, THM desirable set, if offered.
4. Black-and-white positive print, approximately 8x10 inches. Produce 5UA, CRI, THM mandatory image set and 5UA, 5DI, CRI, THM desirable set, if offered.
5. Color lithographed print or card, approximately 3x5 inches or less. Produce 3UA, CRI, THM mandatory image set and 3UA, 3DI, CRI, THM desirable set, if offered.

LOT 2

Sets of digital images for four (4) pictorial items:

1. Representing the type of film in the Mead/Bateson project: Two (2) film frames, each of which is on a multi-frame strip of 35mm film. Each film image shall be reproduced as a 5UA, CRI, THM mandatory image set and a 5UA, 5DI, CRI, THM desirable set, if offered
2. Representing the type of fiche in the Civil War map project: Two (2) 105m fiche, each reproducing one map. Each map shall be reproduced as a CUA mandatory image.

M.2.2.2.1 Evaluation of sample pictorial images

The sample images for both LOT 1 and LOT 2 will be evaluated on a pass/fail basis in terms of the following considerations or features:

- Accuracy of filenames; specifications will be sent with the sample films.
- Files must open and/or decompress in two different IBM-compatible computers, using the following software: PhotoShop and ThumbsPlus.
- TIFF header and tag elements must meet specifications.
- JPEG files will be checked using in JPEGINFO software; header must include JFIF file format indicator.

- JPEG files must be compressed to yield an approximate average reduction of 10:1 for black-and-white images and 15:1 for color images.
- Spatial resolution must be in the range specified in Section C.4.1
- Tonal resolution and tonal quality for images (including check of histogram) must meet requirements in Section C.4.4.1
- Corner to corner sharpness -- judged to the degree possible given image content and quality of original film image
- Cropping -- in terms of the guidelines outlined in Section C.4.8 and in Section J attachments.
- Concatenation -- for the image of the map from fiche only (Lot 2) in terms of the guidelines outlined in Section J, Attachment 6.

M.2.3 Cost

Reasonableness of cost.

M.2.4 Desirable Image Types

After technical evaluation and benchmark testing, those offerors technically acceptable and who have demonstrated the additional capability of producing the desirable image types as specified in C.4.2 may be awarded plus (additional) points. The cost for the desirable images will also be considered.

M.3 52.215-34 EVALUATION OF OFFERS FOR MULTIPLE AWARDS

M.3 52.217-5 EVALUATION OF OPTIONS

Appendix D. Specification for Quality Control and Metadata Building Tool for Managing Scanning Projects

The Quality Control Metadata Building Tool (QCMBT) is intended for use during the physical scanning phase of a digital imaging project to aid the Army Corps of Engineers personnel and/or their contractors to record and review the progress of the scanning process, and as part of the quality assurance program outlined previously in this document. Its reporting capabilities can be used subsequent to the imaging process to identify a document or series of documents by querying the metadata assigned during the imaging process. Its collection of metadata may also be exported for input into an EDMS system as the digital documents created during the imaging process are migrated into a permanent repository.

The specification is now found in a separate document titled “Functional Specification: Quality Control & Metadata Building Tool for Managing Scanning Projects”, ERDC/ITL CR-04-xx.