

**Preserving Cornell's Digital Image Collections:
Implementing an Archival Strategy**

Final Project Report

Anne R. Kenney, Principal Investigator
Oya Y. Rieger, Project Coordinator
Richard Entlich, Digital Projects Librarian

Diane Hillmann
Peter Hoyt
George Kozak
Tim Lynch (advisor)
Tom Turner
Lynne Personius

Cornell University Library

May 2001

Table of Contents

<i>Project Summary and Key Accomplishments</i>	4
<i>Introduction</i>	4
<i>Project Goals</i>	5
<i>Digital Image Collections Inventory</i>	6
<i>File Format Investigation</i>	8
<i>Metadata</i>	9
<i>MARC Clean-Up Project</i>	11
<i>On-the-Fly Conversion</i>	11
<i>Storage Requirements</i>	12
<i>Resource Requirements</i>	13
<i>Preservation Policy Development</i>	14
<i>Dissemination of Project Findings</i>	15
<i>Appendix A: Case Study for Image File Format</i>	16
<i>Appendix B: On-the-Fly Conversion: TIFF2PNG vs. TIFF2GIF Conversion Utilities</i>	22
<i>Appendix C: Expenses Incurred in Creating and Maintaining the Making of America I Collection, 1995-2000</i>	25
<i>Appendix D: Report of the Digital Preservation Policy Working Group</i>	26
INTRODUCTION	29
REQUIREMENTS FOR DEPOSIT	30
1. <i>Selection and Content Considerations</i>	30
A. <i>Scope</i>	30
B. <i>Content/Functional Criteria</i>	31
C. <i>Priorities for Deposit</i>	31
2. <i>Legal Considerations</i>	32
3. <i>Technical Requirements for Conversion</i>	32
A. <i>Source Material for Digitization</i>	33
B. <i>Technical Considerations for Image Files</i>	34
C. <i>Quality Control</i>	38
4. <i>Pre-Depository Storage and Maintenance Requirements</i>	39
5. <i>Metadata Requirements</i>	40
A. <i>Descriptive Metadata</i>	40
B. <i>Structural Metadata</i>	44
C. <i>Preservation Metadata</i>	46
ROLE AND RESPONSIBILITIES OF A CENTRAL DEPOSITORY	51
NEXT STEPS	52
A. <i>Needs Assessment Survey</i>	52
B. <i>Schedule and Procedures for Updating the Deposit Guidelines</i>	53
C. <i>Publicity and Training</i>	53

<i>Appendix 1: Image Quality Assessment</i>	54
<i>Appendix 2: Descriptive Metadata Example</i>	57
<i>Appendix 3: Sample Entry for Digital Image Collections Inventory</i>	59

Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy Final Project Report

Project Summary and Key Accomplishments

The goal of Cornell's IMLS-funded project was to plan and implement an archiving strategy for Cornell Library's digital image collections, representing over 2.5 million images—a half of a terabyte of information. Accomplishments include an inventory of the collections created over the past decade, an investigation of current and emerging file formats for long-term utility, a study of functional requirements for storage, and draft recommendations for preservation metadata. This project also assessed resource needs in terms of staff, equipment, space, time, and finances for a ten-year maintenance program. The project met all of its intended goals, but the main accomplishment had not been anticipated when the grant was submitted. Developing a preservation strategy for the legacy digital image collections convinced project staff that the major challenge was to develop policies and procedures governing *prospective* digital imaging collections that will be mainstreamed into Cornell's Digital Library. With the strong endorsement of the Library Management Team, project staff assembled a library-wide committee to develop a proposal for addressing this problem. Their key recommendation was the establishment of a centralized depository within the Library's Digital Library and Information Technology (D-LIT) infrastructure for ensuring a cost-effective preservation strategy over time. Their report, *Central Deposit Guidelines for Digital Image Collections* (Appendix D), outlines the responsibilities of a *transferee* in preparing an image collection that will be centrally deposited. It also outlines the role and major responsibilities of a central depository, the particulars of which will comprise the second part of this report. The Library Management Team has approved this report and is fully supportive of the second phase effort. We anticipate having in place a fully developed central depository by the end of 2001. The recommendations for digital image creation and metadata contained in this report have also served to define requirements for the Digital Library Federation's proposed digital preservation master files to be included in a DLF-sponsored registry of digitized books and journals.¹ This research is a critical byproduct of the IMLS project that will extend beyond the project deadline and beyond Cornell University.

Introduction

Since 1989, Cornell University Library has been a pioneer in the use of digital image technology to convert library and archival materials to digital form to serve both preservation reformatting and access goals. A primary emphasis of this effort has been the development of quality requirements to support the creation of replacement copies for text-based material. In a series of brittle books digitization projects in the 1990s, printed

¹ "Draft report of a meeting held on 11 April 2001 to consider the potential uses of a service that registers digitized books and journals and to consider implementation."
www.clir.org/diglib/collections/reg/regsum.htm.

pages produced from digital images were compared to the quality of preservation photocopies and computer output microfilm was judged against national standards for preservation microfilming. The conclusion of both these investigations was that 600 dpi 1-bit scanning could result in the creation of replacements for text-based brittle books that was comparable to quality obtained via conventional reformatting means. Throughout the 1990s, Cornell Library followed a hybrid approach in the use of digital image technology, which coupled the creation of digital surrogates for brittle books to improve access with the production of either COM or paper replacements printed on acid free paper for preservation. This approach preserved the informational content contained in the original materials, but did not address the need to safeguard the digital content itself. In using digital imaging, Cornell did not decrease its overall preservation problem but transferred the concern to a new arena.

Because most of the imaging efforts were project-based, by the end of the 1990s, Cornell University Library had assembled an impressive mass of digital information, but it was not managing these digital assets in a manner to ensure their long-term viability. The IMLS-funded project offered the means to plan and implement a coherent digital preservation strategy for Cornell's digital image collections.

Project Goals

Prior to the preparation of the central deposit guidelines, the project team completed a thorough investigation of various technical, administrative, and economic issues related to creating and preserving digital image collections. The key project deliverables include:

- Preparation of an inventory and documentation for Cornell University Library's digital image collections.
- Assessment of current and emerging file formats and compression schemes that are suitable for image collections.
- Investigation of various archival strategies and cost models, and an estimation of resource implications for a selected image collection.
- Placement of the master image files at the center of the digital library system with on-the-fly delivery of access images to support current and future user needs.
- Development and application of descriptive, structural, and preservation metadata requirements.
- Analysis of storage options to assess best approaches to amalgamate, manage, access, and preserve the image collections.
- Identification of key elements of an overall preservation policy for short- and long-term management of the Library's image collection.

These project activities are further described in the following sections. Fuller reports are located on the Project Website (see Figure 1).

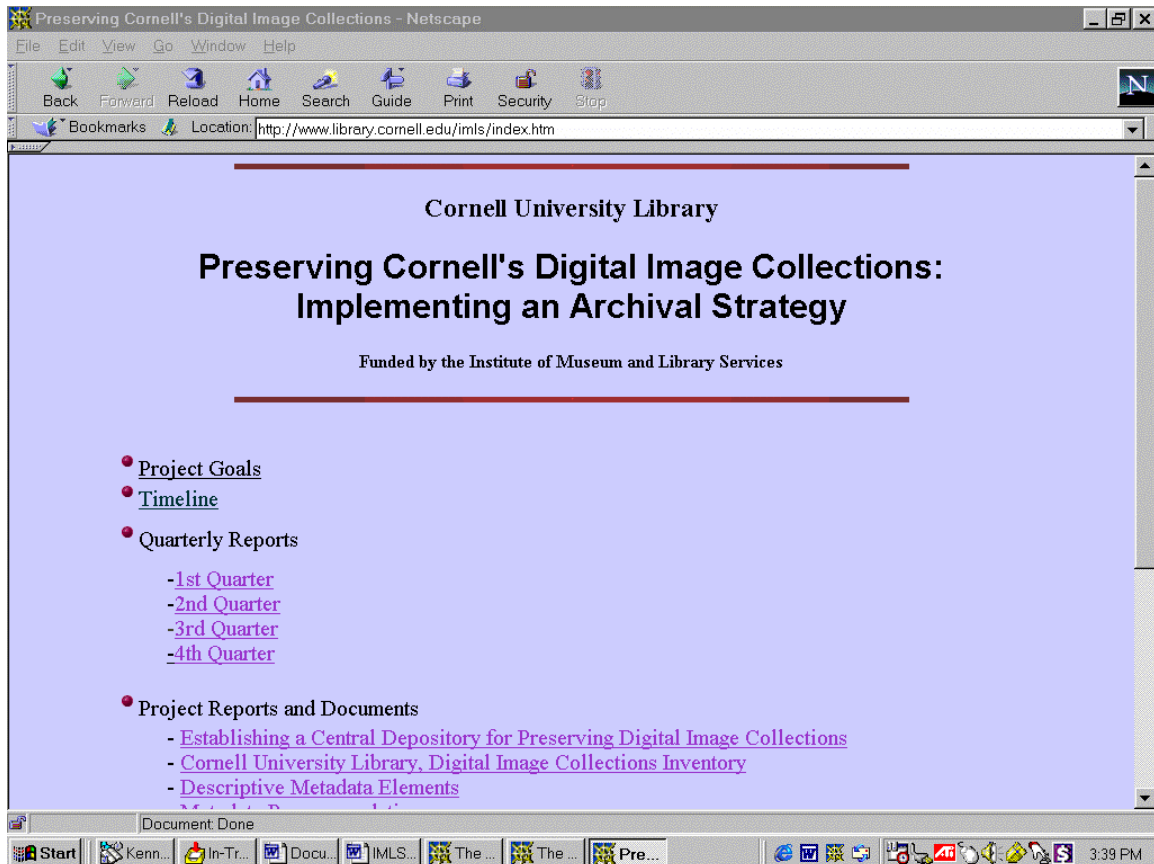


Figure 1: Project Web site: www.library.cornell.edu/imls/index.htm

Digital Image Collections Inventory

One of the first tasks of the project team was to gather information about the Library's existing image collections. An inventory format was developed to gather information such as collection size, conversion specifications, processing information, metadata, access mechanisms, and hardware and software dependencies. The inventory documents the essential descriptive, administrative, and structural metadata for Cornell's image collections (see "Cornell University Library, Digital Image Collections Inventory" on the project Web site). It also demonstrates the difficulties of obtaining this kind of information after-the-fact. Currently, this type of information is not required of staff members who are involved in different stages of digital imaging projects. To avoid such difficulties in the future, the team drafted an inventory form to be completed at the beginning of any project, and recommended the establishment of a Digital Image Collections Inventory database. It aims to include general information that will support preservation administration and decision-making. Table 1 lists the recommended data elements for such an inventory database.

Table 1: Digital Image Collections Inventory: Data Fields

<i>Project description</i>
project title
year the collection was created
project leaders/coordinators, team members
project partners
sources of funding
reason for the project
<i>Source type and characteristics</i>
document type (e.g., printed text, book illustrations, color photographic prints, manuscripts, etc.)
physical dimensions (category: regular, oversize - if possible exact size, or "size varies" statement with min and max measurements -, size varies, but no greater than 8.5 x 11 or some such)
scanned from original or film intermediate
subject matter
<i>Collection size</i>
total file size of the collection including image and metadata files, programs, scripts, etc. (estimated or actual)
number of images
<i>Storage media</i>
type and location
<i>Scanning information</i>
resolution
bit depth
color space or CLUT information for color documents
file format and version
compression technique, version, and ratio
scanner used
vendor vs. in-house scanning
<i>Processing information</i>
any image enhancements on the master copy? E.g., how were halftones handled? Any special treatment?
derivatives created (access, processing; such as scaled/reformatted copies for Web delivery, OCR'ed images, etc.)
<i>Metadata</i>
file header (if possible tags used)
what kind of descriptive metadata – where and how recorded? (e.g., MARC, Dublin, PURL, etc.)
what kind of structural metadata – where and how recorded? (SGML, XML, structuring tags, external metadata, etc.)
what kind of technical metadata – where and how recorded?
special collections – finding aid information
<i>Access mechanisms</i>
online/offline
Web address
<i>System/interface design and characteristics</i>
system specifications (e.g., based on Hunter, Open Text, etc.)
known system requirements
key interface features (forms and style sheets, use of JavaScripts, etc.)
<i>Refreshing/migration history</i>
<i>Rights management & Authenticity</i>
document the process of clearing copyright issues
license information
display and transmission restrictions, right holders
any security/authenticity measures (e.g., watermark)
chain of custody

The short-term plan to implement this inventory is through the development of a Web form, with an automated recording system that can be completed within an estimated time of one hour). In the long-term the plan is to create a DTD for XML (or an XML schema) implementation. Some of the required information is dynamic and cumulative (e.g., refreshing and migration history) and would require updates. The suggested periodic frequency for the revision of the dynamic fields is one year.

File Format Investigation

The project team investigated the current and emerging file formats for long-term utility, including TIFF, JPEG, GIF, PDF, Flashpix, PNG, SPIFF, UPF, SGML/XML, and continues to monitor the development of JPEG2000. After reviewing these file and compression formats, the project team identified the format characteristics for long-term utility, as presented in Table 2.

Table 2: Characteristics of File Formats For Long-Term Utility

General Recommendations:

- Thorough, nonproprietary development and documentation.
- Proven backward compatibility and reliability.
- Wide adoption by large consortia and groups, to increase the chances for well-defined migration paths.
- Support for exchange standards such as the Electronic Document Interchange Standard.

Technical Requirements (Image Quality):

- Support for bit-depth greater than 24 for continuous-tone documents.
- Various lossless and lossy compression schemes.
- Support for various color models and color management features, such as gamma/white point and ICC profiling).
- Multiresolution capability to support both master and access copies.

Metadata:

- Features to encompass/encapsulate various metadata.
- Support for relationship among various components (images, metadata files, scripts, etc.).
- A flexible architecture for metadata recording (e.g., file headers, links to external files, etc.).

Hardware/Software:

- Platform-independent viewing/retrieval software (or cross-platform consistency).
- Minimal hardware/software dependencies.
- Abundant retrieval and image-processing programs for several platforms.

Security:

- Features to report data corruption or error detection and correction.
- Features for authentication.

There is no universal standard format for digital images similar to ASCII for raw text or SGML/XML for encoded text, so project staff recommended the use standard and system-independent file formats, with TIFF serving as the de facto format of choice. One of the outputs of the file format exploration was the creation of a detailed matrix to compare and evaluate the functionality of various file formats and compression schemes, as presented in Table 3, on the following page.

Also included in this investigation was a comparison of different versions of TIFF (4.0, 5.0, and 6.0) and staff explored the potential changes that might be introduced by TIFF 7.0. After a thorough examination of documentation and format tags—and migration tests of current TIFF formats—it was determined that there were no major differences between the versions and that risks involved in migration outweighed the value of converting all files to TIFF 6.0. Staff also investigated issues introduced by storing structural metadata in the proprietary Xerox RDO format and the risks involved in converting to an open format. See Appendix A for a full description of the Image File Format Case Study (also available on the project Web page.)

Metadata

Staff assessed various proposals focusing on the descriptive and administrative metadata that facilitate preservation decisions.² The metadata research included:

- Identifying requisite metadata elements for long-term preservation, including descriptive, administrative, and structuring metadata for serials and monographs (see the Project Web site, "Metadata Recommendations").
- Exploring how and where to record various metadata (descriptive, administrative, structural).
- Evaluating and refining the requisite metadata recommendations and investigated the best ways to mainstream the identification, capture, and maintenance of this metadata for Cornell's existing and new digital collections. A decision was made to expand and apply the *descriptive metadata* proposal first, as implementation of *structural metadata* elements will necessitate resolving some system-related issues first. The library is collaborating with Endeavor to prototype a system (ENCompass) that will provide a sophisticated architecture to house the library's digital image collection. Oya Rieger is participating in a working group on technical metadata for this effort.
- Establishing a repository and naming service (OCLC's Persistent URL [PURL]) and creating persistent identifiers for the image files. Information on Cornell's PURL implementation can be found at: www.library.cornell.edu/voyager/Bibs/ECat/e-cat6.html#6.1

² Metadata recommendations from the following were reviewed: OAIS, National Library of Australia Metadata Proposal, PANDORA, NISO/CLIR/RLG Workshop on Technical Metadata, LC/CNRI Metadata Set, SagaNet, CLIR Hybrid Approach Report, Goettingen Digitization Center, NDLP, MOA1, MOA2, Dublin Core, MARC 007, RLG Metadata Recommendations).

Table 3: Common Image File Formats

Name and Version	TIFF 6.0 (Tagged Image File Format)	GIF 89a (Graphics Interchange Format)	JPEG (Joint Photographic Expert Group)/JFIF (JPEG File Interchange Format)	Flashpix 1.0.2	ImagePac, Photo CD	PNG 1.2(Portable Network Graphics)	PDF 1.3 (Portable Document Format)
Extension(s)	.tif, .tiff	.gif	.jpeg, .jpg, .jif, .jiff	.fpx	.pcd	.png	.pdf
Bit-depth(s)	1 bit (bitonal), 4 or 8 bit grayscale or palette color, up to 64 bit color	1-8 bit (bitonal, grayscale, or color)	8 bit grayscale, 24 bit color	8 bit grayscale, 24 bit color	24 bit color	1-48 bit, 8-bit color, 16-bit grayscale, 48-bit color	4 bit gray, 8 bit color, up to 64 bit color support
Compression	Uncompressed Lossless: ITU G4, LZW, etc. Lossy: JPEG	Lossless: LZW	Lossy: JPEG Lossless: JPEG-LS (not finalized)	Uncompressed Lossy: JPEG	Lossy: "Visually lossless" Kodak proprietary format	Lossless: deflate, an LZ77 derivative	Uncompressed Lossless: ITU G4, LZW Lossy: JPEG
Standard/Proprietary	<i>de facto</i> standard	<i>de facto</i> standard	JPEG: ISO 10918-1/2 JFIF: <i>de facto</i> standard	Publicly available specification	Proprietary	ISO 15948 (anticipated)	<i>de facto</i> standard
Color Mgmt.	RGB, Palette, Y _c bC _r , CMYK, CIE-L*a*b*	Palette	Y _c bC _r .	PhotoYCC and NIF RGB, ICC (optional)	PhotoYCC	Palette, sRGB, ICC	RGB, Y _c bC _r , CMYK
Web Support	Plug-in or external application	Native since Internet Explorer 3, Navigator 2	Native since Internet Explorer 2, Navigator 2	Plug-in	Java applet or external application	Native since Internet Explorer 4 Navigator 4.04, (still incomplete)	Plug-in or external application
Metadata support	Basic set of labeled tags	Free-text comment field	Free-text comment field	Extensive set of labeled tags	Through external databases. No inherent metadata.	Basic set of labeled tags plus user-defined tags.	Basic set of labeled tags.
Comments	Supports multiple images/file	May be replaced by PNG; Interlacing and transparency support by most Web browsers	Progressive JPEG widely supported by Web browsers	Provides multiple resolutions of each image; Wide industry support, but limited current applications	Provides 5 or 6 different resolutions of each image; Unclear future	May replace GIF	Preferred for printing and viewing multi-page documents; Strong government use
Home page	Unofficial: http://home.earthlink.net/~ritter/tiff/	Specification: http://cica.cica.indiana.edu/graphics/image_specs/gif.89.format.txt	http://www.jpeg.org/public/jpeghomepage.htm	http://www.digitalimaging.org	http://www.kodak.com:80/US/en/digital/products/photoCD.shtml	http://www.libpng.org/pub/png/	http://partners.adobe.com/asn/developer/technotes/acrobat.pdf.html

- Standardizing technical metadata through participation in the NISO Technical Metadata for Digital Still Images Committee (<http://www.niso.org/commitau.html>). Project Coordinator, Oya Rieger, is co-chairing this important group, which is expected to release its final report by August 2001. Standardization of technical metadata will facilitate a systematic approach in recording and managing technical image information.

MARC Clean-Up Project

The Technical Services Support Unit and Mann Library Technical Services undertook the clean up of MARC records for two large digital image collections—the Making of America (MOA) and the Core Historical Literature of Agriculture (CHLA). MARC records for the MOA materials were created when the first version of the system was made available in 1996. MARC records for CHLA titles were drafted when that system was being developed. Unfortunately, the cataloging of those records was outdated because standards for cataloging electronic resources and reproductions have changed considerably in the past five years. The updated MARC records can serve as the basis for descriptive metadata to be used for long-term access to these digital materials. Among the changes in the MARC clean-up project were:

- Creating and adding PURLs to the records
- Verifying that access points were properly coded for automated transfer of descriptive elements from MARC to the DC-based structure
- Modifying` MARC record formats to reflect current standards for cataloging
- Improving notes present in the record so they accurately reflect the history of the digitization projects.

Student employees were hired and trained and permanent employees were assigned to work on this standardization project from January 2000 through September 2000. Using resources in this way has put Cornell University Library in a position to take advantage of developments with Endeavor Information System's ENCompass product as well as to describe accurately and effectively these digitized collections.

On-the-Fly Conversion

After holding several discussion sessions to identify on-the-fly conversion requirements for Cornell's digital image collection, a decision was made to focus the efforts on two areas:

- (1) Evaluate and revamp the existing TIFF2GIF conversion mechanism;
- (2) Explore the availability and the utility of TIFF2PNG conversion software and compare its use to TIFF2GIF.

The team evaluated the currently used TIFF2GIF conversion program and adjusted its settings to optimize image quality for access images. The team also examined the potential of replacing GIF with the PNG file format in image delivery, and the pros and

cons of shifting to a TIFF2PNG conversion utility. After identifying and testing conversion software for creating on-the-fly access images in PNG file format to compare the functionality and image quality between TIFF2GIF and TIFF2PNG software, the team wrote a report to summarize the findings. The report is available on the project Web page and is also included as Appendix B.

Storage Requirements

One of the key accomplishments in this area was the exploration of the functional requirements for storage, and to assess alternative solutions for amalgamating, managing, and preserving Cornell's digital image collections. The major focus was not on the media to be used but on the serving capacity of the system for amalgamating all digital image files into one system, simplifying the management of the various collections, maintaining/protecting the digital masters in a live environment, and supporting on-the-fly conversion from the master image files in a secure, timely, and cost effective manner.

The first process was to determine space and storage requirements. The project team determined that the images in question would come to nearly half of a terabyte of 600dpi TIFF images, not including any static derivatives. Since it was the intention of the project team to derive the images for display on-the-fly, the team felt that fast response was needed with an almost immediate access to the archived TIFF version. The system had to be robust and powerful enough to support up to 50 simultaneous user sessions, including searching the image database, processing on the fly, and delivering digital content to the library patron. This eliminated Quantum DLT Tape Libraries or other tape libraries. The team considered Optical Disk Libraries, but decided against them based on problems experienced in the past with these devices. The obvious choice was some form of hard disk storage, possible rack mounted and definitely expandable.

As the team looked at storage solutions (such as IBM SAP Optimized Storage), it became apparent that the best solution was to work with Sun Microsystems, since much of the expertise in the library and at Cornell Information Technologies was with Sun and its equipment and software. A proposal was written to Sun Microsystems for matching funds in kind of equipment to address these issues. The following equipment was obtained from Sun:

- Enterprise 3500 Server with tower enclosure running 2 336MHz/4-MB UltraSPARC modules
- 2GB Memory
- 100MB Ethernet
- 2 internal 9.1GB hot swap disk drives
- Sun Storage A3500 with expansion cabinet originally with 540GB of storage
- SCSI adapter to allow the addition of external drives from our older digital library server as needed

The disks were organized in a RAID-5 arrangement. In RAID-5, independent data disks exist with distributed data blocks so that as each entire data block is written on a data disk, the parity for blocks in the same rank is generated on Writes, recorded in a

distributed location and checked on Reads. In this way, if a disk fails, another disk can be swapped in quickly to fix the problem without loss of data. The ratio of usage is that 1/3 of the disks are used up by the RAID process and not available to the user for storage. This storage was then organized to hold the different collections of digital material that the library had. Software was installed including PERL, the Apache WebServer, and C Compiler to prepare the new system to be the centralized repository for the digital image collections.

To supplement the RAID-5 arrangement, the team also determined the backup requirements for the SUN server and investigated DLT tape drive, ADSM, and Sun Microsystems's Solstice Backup. Cornell Information Technology's ADSM, an IBM-based backup solution, was selected as it allows the customization of services for static directories. Also, the backup is run automatically and handled by a professional staff on a 7x24 basis. The team deemed this as a better use of resources than purchasing a tape drive and doing backups since the cost of the tape drive, media, storage of media and the cost of performing and monitoring the backups on a human resource level was more than the cost of purchasing this service from Cornell Information Technologies.

The RDO³ files, which were created by the proprietary XDOD system, were exported to Cornell Digital Library format. This migration process was essential to be able to amalgamate the collections in the SUN system.

Resource Requirements

After a thorough analysis of existing cost studies for digital preservation, the group decided that there were no existing models that could be readily implemented for our purposes. The main challenge was that most of the existing studies are not itemized and do not indicate what is included in the cost estimates. In addition, most of them present preservation costs on a per gigabyte basis, focusing mostly on storage costs. There are considerable economies of scale in large archives, so calculating costs based on gigabyte units may not be accurate as a system continues to grow. Expenses for digital preservation start accumulating soon after selection for digitization and continue as long as we continue to provide access to the material. To gauge expenses involved in digital preservation, one needs to take into consideration a number of stages in the development of a collection. As expressed in Appendix C, the costs behind long-term management continue to evolve as a collection continues to grow and change.

The team completed the digital preservation resource implications study, with the caveat that for the estimates to be reliable prognostications, such an analysis needs to be conducted after clearly defining the role and responsibilities of the central depository. Based on the findings of that study, the resource implications inquiry will be revisited to identify the staff, equipment, storage space, and funding needs to support ongoing preservation. In preparation for this study, the project team investigated the costs

³ An RDO file contains information about the structure of an image document, as well as a file location pointer for each page image in that document.

involved in developing and maintaining the MOA1 collection during the last 6 years. The findings are presented in Appendix C.

Preservation Policy Development

In the early stages of the investigation of technical issues related to digital preservation, it quickly became evident that managerial and policy issues required equal attention in order to successfully move from a project to program mode. At the end of the first year of the project, the team recommended the development of a strategic plan to ensure the long-term maintenance of the retrospective and prospective digital image collections to the Library's Library Management Team (LMT). The goal behind this proposal was to develop a plan to mainstream digital preservation policies and procedures throughout the Cornell University Library (CUL). This was essential so that the achievements of this project will have a long-lasting value for the CUL, rather than being a one-time effort.

The Digital Preservation Policy Working Group was appointed to develop a mainstreaming strategy for ensuring the longevity of Cornell's digital image collection. The goal was to create a permanent infrastructure to ensure that the concern for the longevity of our image collection will continue beyond the completion of this IMLS-funded project. The Group represented different functional units of the library and aimed to develop a plan to mainstream the implementation of digital preservation policies and procedures throughout the CUL, which is composed of 19 campus libraries. Building a consensus among Cornell colleagues was essential to ensure that the achievements of this project will have a long-lasting value for the library, rather than being a one-time effort. Co-chaired by Anne Kenney and Oya Rieger, the Group completed the central deposit guidelines for Cornell's digital image materials in December 2000. The objectives of the group included:

1. Develop requirements for deposit of digital imaging materials
2. Plan a long-term maintenance and preservation strategy
3. Identify resource requirements to support effective management and preservation
4. Devise a strategy to promote the adoption and consistent use of the baseline standards throughout Cornell University Library

The Digital Preservation Policy Working Group synthesized the findings of the library's IMLS research in a pragmatic policy manual to guide future digital preservation activities. The LMT approved the project team's "Report of the Digital Preservation Policy Working Group: Establishing a Central Depository for Preserving Digital Image Collections PART 1: Responsibilities of Transferee March 2001" (reproduced in Appendix D and also available on the Project Web site). This report outlines requirements for those wishing to transfer digital image resources to the central depository, and covers selection criteria, legal issues, technical imaging requirements, metadata, and storage. The second part of this report will detail the role and responsibilities of the central depository staff. LMT approved the request of the Principal Investigator to appoint a transition team to implement the recommendations on imaging,

metadata creation, storage, and staffing areas and to develop a detailed assessment for establishing the central depository. Work on this transition has been delayed until September 2001 when three key vacancies will be filled: Director of Library Systems, Coordinator of the Digital Imaging and Preservation Research Unit (Oya Rieger assumed a new position in CUL in April), and a programmer who will support the archival repository development.

Dissemination of Project Findings

All project reports have been posted to the IMLS Project Website and the final report will be posted there as well, in both HTML and PDF versions. We consider the principal accomplishment of this project to be the Report of the Digital Preservation Policy Working Group, and the report's availability was announced in the April 15 issue of *RLG DigiNews*, which the Principal Investigator edits ([ww.rlg.org/preserv/diginews/](http://www.rlg.org/preserv/diginews/)). In addition, the report was shared with preservation officers at many research libraries and formed the basis for the Digital Library Federation's recommendations for the creation of Preservation Digital Masters to be included in a proposed Registry of Digitized Books and Journals.

The findings from this IMLS-funded project have also been shared through papers presented by Anne Kenney at the IMLS-sponsored "Web-Wise: A conference on Libraries and Museums in the Digital World," held in March 2000 and by Oya Rieger at the Library and Information Technology Association Forum, in November 1999 and at the MetaE Conference in Alicante, Spain in January 2001. The first presentation was turned into an article "Preserving Digital Assets: Cornell's Digital Image Collection Project," which was published in the June 5th 2000 issue of *First Monday* (http://firstmonday.org/issues/issue5_6/index.html). In addition, project findings have been incorporated Cornell's latest monograph, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, Anne R. Kenney, and Oya Y. Rieger, (Mountain View, CA: Research Libraries Group, Inc., 2000) which won the 2001 Society of American Archivists' prestigious Waldo Gifford Leland Award for "writing of superior excellence and usefulness."). The findings also found their way into Cornell's intensive weeklong digital imaging workshop (www.library.cornell.edu/preservation/workshop/), as well as into its online digital imaging tutorial, available in both English and Spanish at www.library.cornell.edu/preservation/tutorial/. The IMLS project also contributed to the Department's receipt of this year's LITA/Library Hi Tech Award for Outstanding Communication for Continuing Education in Library and Information Technology for its "outstanding contributions in research, continuing education and information sharing to increase the ability of libraries to preserve information for future generations." Finally, the IMLS-sponsored project has prepared Cornell to submit a proposal to the National Endowment for the Humanities in June 2001 to develop and present a new workshop series, entitled "Digital Preservation Management: Short-Term Solutions for Long-Term Problems."

Appendix A: Case Study for Image File Format

1. Collection and Analysis of Source and Target File Format Related Information Investigation Test Bed

To assess the risks associated with file format migration for digital image collections, the project team selected one of Cornell University Library's digital image collections as a test bed. The Ezra Cornell Papers consist of correspondence, financial and legal records, court proceedings, and other documents pertaining principally to the Cornell family, the telegraph industry, and the founding of Cornell University. The collection is composed of 30,000 images stored on small computer system interface (SCSI) disks. They are scanned as 600 dpi, 1-bit TIFF 5.0 ITU Group 4 images. Tag(ged) Image File Format (TIFF) is one of the most popular raster image file formats and is often the format of choice for master image files. It is platform-independent and supports 1- to 24-bit imaging using a variety of compression methods.

The Ezra Cornell materials were scanned in-house using a Xerox scanning system. This system organizes and stores the structuring information (e.g., page number, folder number) in a format called Raster Document Object (RDO), which is Xerox's adoption of the International Office Document Architecture (ODA) and Interchange Format. (1)

Goals of the File Format Migration Investigation

The goals of the file format migration investigation for image files were to:

- identify the TIFF file format attributes at risk during migration
- assess the need to move these TIFF 5.0 image files to the current version (6.0)
- evaluate the risks involved in converting TIFF 5.0 files to TIFF 6.0 files
- investigate the status of upcoming revision to TIFF (7.0)
- assess the risks involved in skipping a generation (TIFF 6.0) and waiting for the release of TIFF 7.0
- assess risks and data loss associated with converting from RDO format to the open Cornell Digital Library (CDL) format.

Collection and Analysis of Source and Target File Format Related Information

To identify digital image format attributes at risk, the project staff collected and analyzed information on different versions of TIFF file format. The research process included the following:

- Conducting a literature search on digital archiving issues pertaining to digital image collections, with a specific focus on migration and the effects of file format choice in the migration chain.
- Investigating new digital preservation research and initiatives, such as the Open Archival Information System (OAIS) reference model, WGBH's Universal Preservation Format (UPF) (Shepard and MacCarn 1999), and the Digital Rosetta Stone Model (Heminger and Robertson 1998), among others.

- Conducting a literature and projects survey to determine the extent of work performed on developing risk analyses based on image files.
- Reviewing risk-assessment tools developed for various purposes, focusing on the form and functionality of these tools and how they can be adapted for the purposes of this project.
- Exploring the dependencies that extend beyond basic image file format attributes, such as internal and external relationships between images and their accompanying metadata files (viewing images as "digital objects" and examining their metadata, associated scripts, programs, etc.).
- Identifying the attributes of digital images that are at risk during format migration, including the effects of migration on metadata, and the various scripts and programs that support retrieval and management of the collection.
- Investigating the existing and emerging bitmap image file formats with a focus on their longevity and other archival attributes.
- Exploring vulnerabilities associated with file format migration and identifying risks associated with "migrating" or "not migrating" these files, with a focus on TIFF files.
- Analyzing the factors involved in decision making in migration projects, such as reformatting a collection of images from TIFF 4.0 to TIFF 5.0 format.
- Examining and comparing the TIFF file format specifications for Versions 4.0, 5.0, and 6.0.
- Exploring the future of TIFF as a file format, with a focus on the characteristics of the TIFF 7.0 format under development.
- Investigating the issues introduced by storing structural metadata in Xerox RDO format.
- Identifying the risks involved in converting RDO files to the CDL format (<http://www2.hunter.com/docs/rfc/rfc1691.html>).

Conclusions of the Source and Target File Format Analysis

Because most of the specifications are publicly available on the Adobe FTP site, the project staff was able to gather a substantial amount of information about the different versions of TIFF. TIFF was developed by Aldus and Microsoft, and the specification was owned by Aldus, which in turn merged with Adobe Systems, Incorporated. Consequently, Adobe now holds the copyright for the TIFF specifications. TIFF is a highly flexible and platform-independent file format. It is supported by numerous image-processing applications. A great strength of the TIFF file format is its file header option, which enables recording within the file itself of a wide variety of metadata (descriptive, administrative, and structural). The set of fields or "tags" in TIFF is extensive, making it the format of choice for most archival reformatting. However, a large number of TIFF fields are not defined by the standard. Therefore, while TIFF offers the advantage of being open and usable, there is the danger that different institutions will define these fields in different ways, leading to problems of compatibility. Another flexibility of TIFF that causes confusion is related to byte order. For example, the TIFF format permits both MSB ("Motorola") and LSB ("Intel") byte order data to be stored, with a header item indicating which order is used.

Tracking the TIFF 7.0 development turned out to be a challenging task. The project team's attempts to contact TIFF 7.0 developers, Adobe, and even TIFF listserv subscribers were fruitless. The TIFF 7.0 development group seems to be determined not to release any information regarding their work. Therefore, the project team was unable to make any comparisons between TIFF 7.0 and the earlier versions. After conducting an extensive evaluation and comparison of TIFF 5.0 and TIFF 6.0 specifications, the team ran several tests to compare the quality and utility of a subset of TIFF 5.0 images before and after conversion to TIFF 6.0. This exploration revealed no major differences between the versions. The project team concluded that there were no risks involved at this point in leaving the testbed images in TIFF 5.0 format. After reaching this conclusion, the team shifted its focus for the risk-assessment study for image files to an examination of storing structural metadata in the proprietary Xerox RDO format. The team will continue to monitor the development of TIFF 7.0.

Raster Document Object Files

An RDO file contains information about the structure of an image document as well as a file location pointer for each page image in that document. A single TIFF file represents each page in the document. The TIFF files each contain the digital data from the scanned page and a header that describes the characteristics of the image file. Because the Xerox Documents on Demand (XDOD) system is proprietary, the structure of image documents can be displayed only by using the appropriate Xerox software.

2. Selection and Evaluation of Conversion Software

Since a decision was made to maintain the files in TIFF 5.0 format, evaluation of the TIFF conversion software was unnecessary. There are several conversion programs on the market for converting TIFF files to various TIFF versions and other file formats (e.g., TIFF to GIF, TIFF to PNG). TIFF 5.0 to TIFF 6.0 conversion could be interpreted as an update rather than as a migration process.

In 1994, Cornell undertook a project to convert the proprietary RDO files to an open CDL format. The specifications for the CDL, which were released in August 1994 through a Request for Comments (#1691), defines an architecture for the storage and retrieval of Cornell University Library's image collection. Similar to RDO files, the CDL document structure provides direct access to the components of image collections (e.g., pages, sections, and chapters).

While the project team's main interest was exploring the export of files created on XDOD 3.0, its immediate concern was with the older RDOs, especially in light of the Y2K compliance issues (i.e., concern that the XDODs would no longer work unless an expensive upgrade was implemented).

The conversion from XDOD RDO to CDL format involved two steps. Cornell used a Xerox-supplied tool (XDOD Export Tool) to convert the RDO files into a series of ASCII

metadata files. This tool is old and can run only in Windows 3.1, and its dissemination is authorized "only pursuant to a valid written license from Xerox." Second, through a locally developed PERL script, the ASCII metadata files were converted to the CDL format. These CDL-formatted structural metadata files are used for navigating through a document (<http://moa.cit.cornell.edu/MOA/EZRA.html>). The Cornell University Library information technology staff wrote the ASCII RDO-to-CDL program.

RDO-to-CDL conversion cannot be achieved through a single software tool since Xerox has not released any RDO specifications.

3. Development of Tools for Assessing the Source-To-Target Format Transfer

No specific software tool was developed to analyze the effects of migration from RDO to CDL format, because all files created using the XDOD scanning system possess identical information fields.

4. Comparison and Analysis after Conversion to Source File Format

The comparison was done manually by comparing the structural metadata elements that were captured in RDO files with the CDL structure. The team compared the list of structural metadata elements captured during scanning with the CDL structuring requirements. All the structural elements mapped to the CDL structure, and there was no loss. Even if there had been a loss, the project team decided that it was much riskier (actually detrimental) to leave the structuring information in an unsupported proprietary format.

5. Releasing the Export Tool to Other Institutions

As part of this project, Cornell investigated the possibility of further developing the Export Tool and making it available to other institutions that have legacy collections in the proprietary Xerox RDO format. This investigation was spurred by two concerns. First, several institutions had requested access to the tool over the past few years, but only Yale University had secured permission from Xerox to use it. Second, in early summer 1999, Xerox informed Cornell that the XDOD 2.x scanning workstations would not be Y2K-compliant without an expensive upgrade. Because Cornell had begun to phase out use of the XDOD systems and had converted all RDO files to the CDL format, our concerns over the millennium focused on our sister institutions' collections.

We initially considered developing the Export Tool into more generic software for external use, but quickly concluded that this would be both expensive and time-consuming. Cornell did not receive any specifications from Xerox for the proprietary tool, and the software developer at Xerox indicated that he doubted that the company still had the tools and specifications to make the system work. We decided to focus on securing permission to release the current version of the Export Tool. After a two-year effort to obtain blanket permission from Xerox to make the tool broadly accessible had

stalled, we turned to documenting the extent of the problem, concluding that Xerox might be more amenable to a very limited release.

In late April 1999, Cornell posted the following announcement on 11 listservs.

Export Tool to Convert Xerox RDO Files to Open Digital Library Format

Has your institution created digital image files using the proprietary Xerox Documents on Demand software that generates Raster Document Objects (RDOs) to store structural metadata? Cornell University is seeking feedback from these institutions to determine what demand there would be for freeware to convert those RDOs for use in other metadata applications. Cornell has used the RDO2CDL export tool to migrate RDOs to ASCII metadata files that recreate the logical and physical structure format of the RDO (called CDL). If your institution is interested in utilizing such an Export Tool, please send contact information and a brief description of your needs to: Anne R. Kenney (ark3@cornell.edu).

By early June, surprisingly few responses were received. Universities with files created on XDOD 2.5 or older versions included Harvard, Penn State, the University of Tennessee-Knoxville, and Yale. Those responding with files created using XDOD 3.01 or DigiPath included the Hein Publishing Company, Illinois State Library, the National Document Center (Athens), Indiana University, the University of Toronto, and the National Oceanic and Atmospheric Administration (NOAA) Miami Regional Library (which was considering using the technology).

Inquiries to Xerox about releasing the tool to this group resulted in further clarification that the RDO Export Tool software would work only as configured on XDOD Version 2.x systems. The format of the RDO changed slightly from version 2 to version 3, and the Export Tool would not convert the structural data on version 3 or higher systems. The Hein Publishing Company had used the tool with version 3 files through a collaborative project with Cornell, but only page labels, not structuring information, were exported. William Anderson, the Xerox software engineer who created the tool, suggested that it would be possible to get the structure information out of the version 3 RDO files, but it would take a programmer with knowledge of the Office Document Architecture (of which RDO is a variant), fair knowledge of Unix tools, and a copy of the RDO Version 3 specification, which Xerox seemed unwilling or unable to make available publicly. Anderson suggested that, "If customers are looking to buy DigiPath today, and they need that facility, they should ask for it." Xerox decided to grant access to this software only to XDOD 2.x customers who were not migrating to DigiPath.

From June to early September, efforts continued to reach legal agreement with Xerox over the release of the Export Tool software to XDOD 2.x users. Cornell received a copy of a proposed Software License Agreement on August 26, 1999. The agreement granted the institution a nonexclusive, perpetual, royalty-free license to use the software and the right to provide a sublicense only to those institutions that had reported using the XDOD Version 2.x systems, collectively referred to as "Authorized Educational Institutions"

(AEI). Lee Cartmill, the chief financial officer at Cornell University Library, expressed concern about the indemnity clause in the agreement, which required Cornell to "defend, indemnify and hold Xerox harmless from and against any and all third party claims that arise from or relate to the Software and their respective use of the Software." Cornell attempted to have this clause modified. When Xerox remained adamant, Cartmill drafted a Software Sublicense Agreement that would require the AEIs to extend the indemnity and limitation of liability to Cornell University. The four institutions were notified of these stipulations, and their legal advisers reviewed copies of the agreements. Only three institutions agreed to the license stipulations, and the software was released to them.

References

International Organization for Standardization. ISO Reference Model for an Open Archival Information System (OAIS). 1998. Available from <http://ssdoo.gsfc.nasa.gov/nost/isoas/us/overview.html> .

Heminger, Alan R., and Steven B. Robertson. 1998. *Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents*. Available from <http://crack.inesc.pt/events/ercim/delos6/papers/rosetta.doc>.

Shepard, Thom, and Dave MacCarn. 1999. UPF: Universal Preservation Format. Available from <http://info.wgbh.org/upf/>.

Endnotes

1. ODA, which became an ISO standard in 1988, has been developed to represent and allow the interchange of office documents. It contains facilities that allow both the structure and content of complex multimedia documents to be represented. Although ODA is an open standard, specifications for the RDO architecture are proprietary.

Appendix B: On-the-Fly Conversion: TIFF2PNG vs. TIFF2GIF Conversion Utilities

Why On-the-Fly Conversion?

Document scanning that emphasizes full, high-fidelity capture has many advantages, including the creation a rich digital master that can serve many needs. One of the drawbacks, however, is the difficulty in delivering such master files via the Internet. High-resolution master files take a long time to transmit and cannot be comfortably viewed, even on state-of-the-art displays, due to their large pixel dimensions. Therefore, a central part of Cornell's digital preservation strategy has been to place the high-quality master image files at the center of the delivery systems and to develop on-the-fly conversion capabilities to create derivatives to meet varying user and staff needs. As strongly advocated by John Price-Wilkin, this approach supports flexibility in delivery, without reliance on the creation of static derivative images that would support only certain time-sensitive and specific needs. This approach also helps to reduce costs as derivatives are created only when there is a request.⁴

On-the-Fly Conversion Software

On-the-fly conversion is an elegant solution to the Web delivery problem associated with high-quality master images. Made feasible in recent years by the wide availability of high-speed workstations, on-the-fly conversion can quickly produce small yet legible versions of master files on an as-needed basis. The technique validates the theory of the rich digital master by allowing it to serve multiple purposes without requiring the long-term storage of derivative files.

On-the-fly conversion has been most commonly used to scale high-resolution, bitonal TIFF files to lower resolution, gray-enhanced GIF files for Web delivery. A fast conversion utility called *tif2gif* was developed by the University of Michigan to carry out TIFF to GIF conversions and allows the selection of the degree of scaling and the number of bits of gray to add. Conceivably a TIFF file could be scaled and gray-enhanced without conversion to GIF, but TIFF is not one of the Web's native graphics formats. GIF, on the other hand, has been widely supported by Web browsers since the early days of the World Wide Web. Though GIF is limited to 8-bit depth, gray enhancement of TIFF files usually employs only three to five bits of gray.

However, despite the absence of any technical limitations in using GIF as the target for on-the-fly conversion of high-resolution bitonal files, there are other good reasons to seek a substitute. The compression algorithm used in GIF files is patented, and the patent holder, Unisys Corp., has been pursuing royalty payments in recent years from software producers that use its algorithm and from web sites that employ GIFs whose origin in

⁴ For a full discussion of the virtues of on-the-fly conversion, see: John Price-Wilkin. Enhancing Access to Digital Image Collections: System Building and Image Processing in *Moving Theory into Practice: Digital Imaging for Libraries and Archives* by Anne R. Kenney, Oya Y. Rieger. Mountain View, CA: Research Libraries Group, Inc., 2000.

licensed software cannot be proven. As a result, the Internet community developed PNG (portable network graphics) a royalty-free graphics file format to take the place of GIF.

PNG is emerging, along with the already established GIF and JPEG, as the third natively-supported graphics format on the Web. As part of this investigation, we researched the status of PNG and reported the results in the FAQ of the August 15, 2000 issue of RLG DigiNews (see <http://www.rlg.org/preserv/diginews/diginews4-4.html#faq>). Given PNG's improving support and a growing movement to retire GIF as an active Web graphics format, it makes sense to examine the available tools for converting high-resolution, bitonal TIFF files to PNGs instead of GIFs.

TIFF2PNG vs. TIFF2GIF

Thus our second task was to conduct a survey of TIFF to PNG conversion tools. We started our survey close to home. We were aware that a Cornell staff member had created a TIFF to PNG conversion utility (*tif2png*) and was using it to create Web graphics for some sites. We were curious how the resulting PNG files would compare to corresponding GIF files created by the University of Michigan *tif2gif* utility. We were interested in parameters such as conversion speed, output quality, flexibility of gray level enhancement, etc.

However, in discussing the nature of the *tif2png* utility with its creator, it became evident that no significant differences should be expected. *Tif2png* was created by taking the program code from *tif2gif* and changing only the portion that converts the TIFF file to another format. The code that does the scaling and the gray-level enhancement is identical. Since GIF and PNG both use lossless compression, there should be no differences in quality.

Further discussions with the program's creator revealed that in his experience, the speed of conversion is very similar, with *tif2gif* slightly faster than *tif2png* on some files and vice versa. Converted GIF and PNG files resulting from the same starting TIFF file were closely compared and found to be identical down to the last bit, as long as the same command line parameters (degree of scaling and number of bits of gray enhancement) are used.

We found one other TIFF to PNG utility, co-authored by one of the designers of the PNG format and available at <http://www.libpng.org/pub/png/apps/tif2png.html>. This utility is available for the DOS and Linux operating environments, but in its present form supports neither scaling nor gray-level enhancement. Therefore, it does not serve as a suitable substitute for *tif2gif*. It is possible that this utility could be part of the toolkit for delivering gray scale and color TIFF files over the web (as converted PNG files), since *tif2gif* only works on bitonal images. However, it would be desirable to incorporate variable scaling to reduce the size and resolution of the images.

During on-the-fly conversion investigation, we discovered that the grayscale settings the tiff2gif utility that Cornell is using was not optimized for the delivery of bitonal images with illustrations. This problem was fixed at once.

The third leg of our investigation was to query the digital library community about its current level of PNG utilization and for any observations about the quality or usability of those images. A message⁵ was sent out on several listservs requesting feedback

One institution reported that upon receiving a request for an image, if its web site detected a PNG-enabled browser, it would deliver the image as a PNG instead of as a GIF. Another institution reported that it had considered using PNG for image delivery and had hired a contractor to run some tests. The images received were judged to be unacceptably distorted for their intended purpose. Given that PNG is a lossless raster image format, there is no reason files converted to PNG show any distortion at all. So the problem reported probably resulted either from the use of improperly written conversion tools or errors on the part of the contractor. Either way, the incident points out some of the difficulties faced by a new format trying to gain acceptance against established formats, even when it offers good functionality and no licensing difficulties.

Conclusion

In summary, our investigation revealed that although PNG is a well-designed replacement for GIF, it currently faces several obstacles which preclude its immediate adoption as a target for delivery of converted, high-resolution, bitonal TIFF files. These obstacles include insufficient web browser support and an overall lack of knowledge and interest on the part of the library imaging community.

The on-the-fly conversion study was limited to bitonal images. We have decided to continue using the tiff2gif conversion software (for delivery of bitonal text-based images) for the time being. However, we will continue our investigation as new file formats and compression techniques continue to emerge. One interesting trend that we are closely watching is the increasing popularity of file formats that provide nested-resolution (a.k.a. tiling formats). This category of file formats (e.g., FlashPix, GridPix, and JTIP, JPEG2000) is particularly appropriate for tonal images, and have the advantage of fast delivery and high quality. Also certain compression techniques such as Wavelet allow panning and zooming to fit the file to the purpose on the fly. With server-side wavelet compression, users can dynamically create JPEG derivatives at various resolutions and focus on specific segments of an image.

⁵ Cornell University Library, Department of Preservation and Conservation is involved in a small project to compare the quality and utility of GIF and PNG file formats in the delivery of access image files. The goal of the study is to compare various attributes of GIF and PNG file formats for on-the-fly delivery purposes (generation of GIF or PNG files for online delivery from master TIF images when there is a request). Although the PNG file format has gotten quite a bit of publicity (and partial Explorer and Navigator support) during the last couple of months, it looks like the file format has not yet become very popular in image delivery on the Web. Please contact me if you know of any institutions or projects that use PNG file format for master or access images.

Appendix C: Expenses Incurred in Creating and Maintaining the Making of America I Collection, 1995-2000

(Note: double click on the spreadsheet to bring it up, with explanatory notes, in MS Excel.)

Category of expense	Years during which expenses were incurred				
	1995-1996	1997	1998	1999	2000
selection & copyright clearance	\$ 5,540.00	\$ -	\$ -	\$ -	\$ -
preparation	\$ 20,700.00	\$ -	\$ -	\$ -	\$ -
benchmarking (decision making)	\$ 5,540.00	\$ 2,000.00	\$ 2,000.00	\$ 2,000.00	\$ 2,000.00
conversion	\$ 126,000.00	\$ 3,000.00	\$ 6,000.00	\$ 3,000.00	\$ -
descriptive metadata & indexing	\$ 83,700.00	\$ 1,385.00	\$ 1,385.00	\$ 15,000.00	\$ 15,000.00
OCR	\$ -	\$ -	\$ -	\$ 14,808.00	\$ 14,808.00
OCR Software/Hardware	\$ -	\$ -	\$ -	\$ 5,200.00	\$ -
SGML & structuring metadata	\$ -	\$ -	\$ -	\$ 11,080.00	\$ 11,080.00
SGML Software/Hardware	\$ -	\$ -	\$ 33,000.00	\$ 5,000.00	\$ -
QC	\$ 15,300.00	\$ 32,000.00	\$ 4,000.00	\$ 7,404.00	\$ 7,404.00
project management	\$ 11,080.00	\$ 5,540.00	\$ 5,540.00	\$ 5,540.00	\$ 2,770.00
derivative creation & post processing	\$ -	\$ 7,404.00	\$ -	\$ -	\$ -
printing	\$ -	\$ 109,000.00	\$ -	\$ -	\$ -
binding		\$ 7,800.00			
CUL storage and maintenance	\$ 12,000.00	\$ 38,390.00	\$ 6,000.00	\$ 60,000.00	\$ 6,000.00
programming/system support	\$ 37,020.00	\$ 14,808.00	\$ 14,808.00	\$ 22,212.00	\$ 22,212.00
interface design	\$ -	\$ 7,404.00	\$ 3,702.00	\$ 7,404.00	\$ 3,702.00
needs assessment & user evaluation	\$ 24,600.00	\$ -	\$ -	\$ 2,770.00	\$ -
users support (reference)		\$ 1,385.00	\$ 2,770.00	\$ 1,385.00	\$ 1,385.00
CIT storage facility	\$ 3,500.00	\$ 3,500.00	\$ 3,500.00	\$ 3,500.00	\$ 3,500.00
networking	\$ 3,000.00	\$ 3,000.00	\$ 3,000.00	\$ 3,000.00	\$ 3,000.00
refreshing	\$ -	\$ -	\$ -	\$ 22,212.00	\$ 7,404.00
migration	\$ -	\$ -	\$ -	\$ 14,808.00	\$ 7,404.00
digital archeology	\$ -	\$ -	\$ 3,000.00	\$ 3,000.00	\$ -
supplies	\$ 6,000.00	\$ 6,000.00	\$ 6,000.00	\$ 6,000.00	\$ 6,000.00
travel & consultancy	\$ 5,108.00	\$ 5,108.00	\$ 5,108.00	\$ 5,108.00	\$ 5,108.00
overhead costs	\$ 2,000.00	\$ 2,000.00	\$ 2,000.00	\$ 2,000.00	\$ 2,000.00
TOTAL	\$ 361,088.00	\$ 249,724.00	\$ 101,813.00	\$ 222,431.00	\$ 120,777.00
INFLATION FACTOR	\$ 10,832.64	\$ -	\$ -	\$ 4,003.76	\$ 4,347.97
GRAND TOTAL	\$1,075,017.37				
<i>number of images</i>	<i>910,000</i>	<i>total file size</i>	<i>175 gigabytes</i>	<i>cost/image/6 yrs</i>	<i>\$ 1.18</i>

Appendix D:

Report of the Digital Preservation Policy Working Group
on
**Establishing a Central Depository for Preserving
Digital Image Collections**

PART 1: Responsibilities of Transferee

Version 1.0, March 2001

Anne R. Kenney and Oya Y. Rieger, Co-Chairs

David Block
Erla Heyns
Peter Hirtle
George Kozak
Joy Paulson
Tom Turner

Cornell University Library

Central Deposit Guidelines for Digital Image Collections

Final Project Report	4
Project Summary and Key Accomplishments	4
Introduction	4
Project Goals	5
<i>Digital Image Collections Inventory</i>	6
Table 1: Digital Image Collections Inventory: Data Fields	7
<i>File Format Investigation</i>	8
<i>Metadata</i>	9
<i>MARC Clean-Up Project</i>	11
<i>On-the Fly Conversion</i>	11
Storage Requirements	12
Resource Requirements	13
Preservation Policy Development	14
Dissemination of Project Findings	15
Appendix A: Case Study for Image File Format	16
Goals of the File Format Migration Investigation	16
Collection and Analysis of Source and Target File Format Related Information	16
Raster Document Object Files	18
Export Tool to Convert Xerox RDO Files to Open Digital Library Format	20
References	21
Endnotes	21
Appendix B: On-the-Fly Conversion: TIFF2PNG vs. TIFF2GIF Conversion Utilities	22
Appendix C: Expenses Incurred in Creating and Maintaining the Making of America I Collection, 1995-2000	25
<i>Appendix D:</i>	26
INTRODUCTION	29
REQUIREMENTS FOR DEPOSIT	30
1. <i>Selection and Content Considerations</i>	30
A. <i>Scope</i>	30
B. <i>Content/Functional Criteria</i>	31
C. <i>Priorities for Deposit</i>	31
2. <i>Legal Considerations</i>	32
3. <i>Technical Requirements for Conversion</i>	32
A. <i>Source Material for Digitization</i>	33
<i>Scanning from Originals vs. Intermediates</i>	33
<i>Master vs. Derivative Files</i>	34
B. <i>Technical Considerations for Image Files</i>	34
Table 1: <i>Digital Master Image Files— Recommended Imaging Requirements</i>	36
Table 2: <i>Digital Master Image Files— Minimal Imaging Requirements</i>	37
C. <i>Quality Control</i>	38
<i>QC Recommendations</i>	38

4. <i>Pre-Depository Storage and Maintenance Requirements</i>	39
5. <i>Metadata Requirements</i>	40
A. <i>Descriptive Metadata</i>	40
B. <i>Structural Metadata</i>	44
Table 3: <i>List of Structural Metadata Elements for Digital Image Collections</i>	45
C. <i>Preservation Metadata</i>	46
Digital Image Collections Inventory	47
Table 4: <i>Digital Image Collections Inventory: Data Fields</i>	48
Technical Metadata	49
Authenticity	50
<i>ROLE AND RESPONSIBILITIES OF A CENTRAL DEPOSITORY</i>	51
<i>NEXT STEPS</i>	52
A. <i>Needs Assessment Survey</i>	52
B. <i>Schedule and Procedures for Updating the Deposit Guidelines</i>	53
C. <i>Publicity and Training</i>	53
<i>Appendix 1: Image Quality Assessment</i>	54
<i>Appendix 2: Descriptive Metadata Example</i>	57
<i>Appendix 3: Sample Entry for Digital Image Collections Inventory</i>	59

INTRODUCTION

This document presents recommendations from the Digital Preservation Policy Working Group, which was charged with developing a prospective strategy for managing Cornell's digital image assets over time. The Working Group itself represented a logical extension of the planning efforts undertaken as part of a project funded by the IMLS to develop a digital preservation solution for Cornell's retrospective digital image collections created over the past decade. The Working Group membership included individuals from across the library system who have the requisite expertise and/or responsibility for selecting, managing, and serving digital image collections. Anne Kenney and Oya Rieger co-chaired the Working Group, and the membership included: David Block and Erla Heyns (selection and content considerations), Joy Paulson (image conversion), Peter Hirtle (legal and technical requirements), George Kozak (technical requirements), and Tom Turner (metadata).

This report begins with two important recommendations to the Library Management Team. The first one is to establish centralized responsibility for ensuring continuing access to digital image collections over time. *The Working Group recommends that this responsibility take the form of a central depository to be administratively located within the Digital Library and Information Technology (D-LIT) infrastructure.* It is the Working Group's strong belief that centralized responsibility will facilitate the long-term use of digital resources in the most cost-effective manner.

Although preservation is the driving force behind this recommendation, in the digital realm, preservation is met by the ability to continue to provide reliable and trusted access. Current forms of access to materials may be managed remotely and/or locally. *The Working Group recommends that the staff of the central depository collaborate with the transferee to provide both access to and security of files entrusted to their care.* The transferee may continue to exercise a principal role in supporting access to the content or assume a consulting role in the event of any significant change in functionality or service. In some cases, such as with text-based images, both preservation and access can be served by placing the master image files at the heart of the delivery system.

With these two recommendations in mind, the Working Group has outlined a plan for establishing a central depository. This report represents the first installment in that plan. It defines requirements for those wishing to transfer materials to the central depository and outlines the role and responsibilities that a central depository might have. The Working Group decided not to outline these latter responsibilities in detail until the Library Management Team approves the concept of a central depository in principal. The Working Group also acknowledges that a fulsome description of the depository should be based on an assessment of the Cornell University Library's readiness to support such a responsibility. This assessment was beyond the scope of the Working Group's charge. The report concludes with a recommendation to conduct a three-month feasibility study, the details of which are outlined in the "Next Steps" section (page 29). Following this feasibility study, the second installment of the plan can be undertaken, which will provide a detailed description of the Central Depository.

REQUIREMENTS FOR DEPOSIT

This section outlines the responsibilities of a *transferee* in preparing an image collection that will be centrally deposited. Long-term management of digital collections requires a substantial commitment of institutional resources. The guidelines aim for a level of homogeneity, with the assumption that the maintenance of collections that are created based on consistent techniques is more practical and cost-effective in the long run.

Digital image collections can be of two types: those created initially via a digital process, and those that represent digital surrogates for analog source documents. Currently the vast majority of digital image files managed by Cornell units come from this latter category. To be eligible for inclusion in the central repository, and in order for them to be managed in an effective and efficient manner, image files must meet some minimal requirements. Therefore, guidance and recommendations for the selection, creation, and storage of digital image files and related documentation are included in this section. Although these recommendations may prove useful to projects of a short-term nature, they are not to be viewed as binding on such projects, nor are they designed to inhibit individual initiatives involving the use of digital image technology. The guidelines will be used only to determine whether to invest in the long-term care of digital collections by accepting them for central deposit.

1. Selection and Content Considerations

This section of the report focuses on the scope and nature of the digitized content that will be selected for the central depository.

A. Scope

Digitized materials, like their analog counterparts, should primarily support teaching and scholarship at Cornell University and secondarily the needs of other communities. Selection of digital image materials is based, in significant part, on their content. It is also essential that the functionality of the digitized items be considered as part of the selection decision, i.e., the extent to which the items' particular digital forms affect or extend their utility for specific scholarly or educational purposes. Additional evaluation elements, enumerated in subsequent sections of these guidelines, imply that a resource meets the standards of authority and relevance that characterize other segments of the library's holdings.⁶

⁶ Half a decade of reflection, deliberation and implementation has expanded selection criteria to include analysis of what digital collections imply for technical processing, physical access, interpretation and preservation. While these considerations await formal codification, their essence appears in scattered documents: Report of the Committee on Electronic Resources (1996) <http://www.library.cornell.edu/DLWG/CERREPOR.htm#Rec>, Report of the Full Text Working Group, www.mannlib.cornell.edu/ftrat/ftwg-rpt.pdf (1998) and the Criteria for Comparative Assessment of Networked Resources, an instrument for assessing electronic databases prior to initial selection or relicensing <http://www.english.cornell.edu/cul/a2i/drc/ccaform.html> [1999] among them.

B. Content/Functional Criteria

Digital image materials selected for central deposit should reflect the same content criteria used for evaluating print collections. They should:

- Support the curriculum, including emerging distributed learning initiatives
- Ensure standard source availability—the identification of “core” publications
- Facilitate faculty and student research
- Maintain national collection strengths
- Honor inter-institutional and other commitments

These considerations are codified in the collection development policy of the Library:

Cornell University Library, Collection Development Policy Statements
www.library.cornell.edu/colldev/cdhome1.html

Cornell Primary Collecting Responsibilities
www.library.cornell.edu/colldev/cpcr.pdf

C. Priorities for Deposit

At some point, CUL may have to make difficult choices about which materials to accept for deposit based on economic or managerial considerations. Technical and legal issues will most likely be the determining factors in selection but content decisions may also come into play. While not restricting selection to materials that meet one or more of the following criteria, preference for central deposit could be given to digital image materials that:

- Represent complete and credible versions of digitized resources
- Represent thematic or format-based aggregates, rather than idiosyncratic works
- Help create a comprehensive collection
- Enhance access to collections by making them easier to browse, search, and use
- Increase use of collections, by bringing little-known materials to light or by widening potential readership
- Are accessible campus-wide
- Help preserve, protect, store remotely, or replace materials by providing reliable surrogates for consultation
- Enable new kinds of research, not possible in the analog form
- Do not duplicate resources available to the Cornell community via other arrangements, e.g., digital resources managed at individual libraries within the system or other divisions within the University, or made accessible via other institutions, consortia, inter-institutional agreements, or existing licenses

2. Legal Considerations

The ready accessibility of digital documents distinguishes them from their print counterparts, and complicates their selection. The digital equivalent of practices that were common with print material may place the University at unacceptable risk merely because they are more visible and may have greater consequences. Selection for central deposit must begin with an understanding of the current uncertainty in the application of copyright to digital resources. Copyright law, still written on and, most cogently about paper, is in flux as it extends to digitized materials.

The uncertainty in how copyright law will be applied to digital materials leads to an additional selection criterion: Cornell's legal right to store digital copies of material can be justified. There are several possible justifications for central management of a digital image resource:

- The digital image materials are in the public domain.
- The copyright holder has granted permission for network distribution and use.
- An assessment of the risk involved in digitizing a collection reveals that it is unlikely Cornell could be found at fault, either because the copyright owners cannot be found or the use is presumed to be fair.
- The digital resources are copies of unpublished works found in a different library.
- The digital copy has been made to replace a damaged, deteriorating, lost, or stolen copy of a work that cannot be obtained at a fair price, and the digital copies of which will not be made available outside of the library.

Evidence of the copyright status and documentation of efforts to obtain permission to make copyrighted digital resources available—including a signed copyright waiver from the copyright holder or written documentation that details a good faith effort to secure such permission—are required as part of the deposit process. Under special circumstances, digitized material that is copyright-protected, but which will fall into public domain within a short time frame, may also be considered for deposit. A chart summarizing the terms of protection for published and unpublished materials is available at <http://cidc.library.cornell.edu/copyright/>. Note that copyright may cover software use as well as digital content. In addition to copyright, privacy and donor restrictions must be considered. It is the depositing unit's responsibility to ensure that these rights are not breached by the digitization and use of such materials.

3. Technical Requirements for Conversion

There are compelling preservation, access, and economic reasons for creating rich digital master image files that reflect all significant informational content contained in the original source materials. These files have the best chance of remaining useful and cost effective over time for a number of reasons. Preservation is one of the main arguments for rich digital masters. Digital files can be created to reduce use of or in some cases replace a deteriorating or vulnerable original, provided the digital surrogate offers an

accurate and trusted representation. Preservation of the digital files themselves is also served when digital images are captured consistently, the capture methods are well documented, and widely supported file formats are used. And it makes good economic sense to produce sufficiently high-level images to avoid the expense of reconverting when technology requires or can use a richer digital file. This point is particularly compelling since the expense of identifying, preparing, inspecting, and indexing digital information far exceeds scanning costs. The master file can also be used to create derivatives files, which meet a variety of current and future users' needs. The quality, utility, and expense of derivatives for publication, image display, and computer processing are directly affected by the quality of the initial scan.

A. Source Material for Digitization

Digital image files can be created to serve as surrogates for a range of document types. Source material can include, but may not be limited to, the following:

- Printed text—distinct edge-based representations that are cleanly produced, with no tonal variation, such as a book containing text and simple line graphics
- Book illustrations—representing the range of relief, intaglio, and planographic illustration processes reproduced in books produced in the 19th and 20th century, including halftones, etchings, and engravings
- Rare or damaged printed text—items that convey intrinsic information beyond the printed text or those in which the text may be obscured by surface dirt, stains, or other damage
- Manuscripts—soft, edge-based representations that are produced by hand or machine, but do not exhibit the distinct edges typical of machine processes, such as a letter or line drawing
- Maps, architectural drawings—oversized materials that contain fine details, line drawings, and text, either hand or machine-produced
- Graphics—original relief, intaglio, and planographic illustrations
- Works of art on paper—hand produced artwork, including water colors, charcoal sketches, pencil drawings, tempura, and oil painting
- Photographic prints—reflection print formats, including cartes de visite, cabinet cards, 3.5" x 5", 4" x 5", postcard, 5" x 7", and 8" x 10"
- Photographic transparencies and negatives—negatives and positive transparencies produced on film or glass, including 35 mm, lantern slides, 4" x 5", 5" x 7", and 8" x 10"
- Microformats, including 16 mm, 35mm, 70 mm microfilm and 105 mm microfiche

Scanning from Originals vs. Intermediates

As a general rule of thumb, scanning from the original will ensure the highest quality image file. When multiple copies of an item exist, scan from the best copy available, whenever possible. The use of an intermediate, such as a slide, transparency, microfilm, or photocopy, will introduce another step in the imaging process, increasing the

complexity of the workflow, and lowering the quality of the resulting image. It may also affect the accuracy of subsequent image processing, such as the use of Optical Character Recognition (OCR) programs. Quality and processing applications will be particularly compromised if the intermediate itself is poorly produced, damaged, or in deteriorated condition. If an intermediate is used, ensure that it has been prepared according to established standards and is in good condition, free of scratches, dust, light damage, and distortions.

Master vs. Derivative Files

These guidelines cover technical requirements in the creation of master images only. Derivative files that are currently being used for access (e.g., thumbnails, access images, printing files) may also be deposited but they will be maintained only as long as they are being used in the access system. There is no guarantee that they will be maintained for any set length of time, as derivative requirements may change rapidly based on developments in file formats, compression techniques, display technologies, and on-the-fly generation capabilities. *Projects considering the development of static derivatives should consult with Depository staff about current practice in creating access versions.* For instance, derivatives of text-based image files currently include: 10 dpi 3-bit GIF thumbnails, 100 dpi and 75 dpi 3-bit GIFs created on the fly, and text files produced via OCR. PDFs of individual pages are created for printing purposes from the master images (e.g., 600-1bit TIFF images).

B. Technical Considerations for Image Files

The following technical considerations are covered in the Recommended Imaging Requirements. Additional technical considerations are discussed under Quality Control.

- Resolution—the spatial frequency at which a digital image is sampled, often stated as dots per inch (dpi), pixels per inch (ppi), or pixel dimensions.
- Bit depth—determined by the number of bits used to define each pixel. Digital images may be produced in black and white (bitonal, or 1 bit images), in grayscale, or in color. All source documents containing intentional color or where discoloration provides important evidence of age and use should be imaged in color.
- Enhancement/image processing—any process applied to the raw scan to improve quality or legibility. Generally accepted enhancements included reduction of greater than 8-bit/channel linear data to 8-bit non linear data; contrast stretching; minimal adjustments for color and tone; descreening/rescreening of halftones and other graphic content to reduce/minimize moiré.
- File format—consists of both the bits that comprise the image and the header information on how to read and interpret the file. Currently there is no clear archival format to recommend, although open, widely supported file formats are recommended, with preference given to TIFF files 5.0 and 6.0. The goal is to limit the number of file formats that need to be managed by the Depository. The future

of TIFF and other formats, however, is uncertain, and the depository staff will need to monitor format development. A table presenting information on some of the more common image formats in use today is available at <http://www.library.cornell.edu/preservation/tutorial/presentation/table7-1.html>.

- Compression—a process used to mathematically reduce or abbreviate the string of binary code in an uncompressed image. Compression techniques can be either lossless (no information discarded in the process) or lossy (where the least significant information is averaged or discarded). There is a clear preference for uncompressed files or for compressed files using lossless compression. The goal is to limit the number of compression processes that need to be managed long term. A table listing attributes of common compression formats is available at <http://www.library.cornell.edu/preservation/tutorial/presentation/table7-3.html>.

Two levels of technical imaging requirements are presented in the charts below. The first represents Recommended Requirements that will promote the long-term viability of digital image collections, and are detailed in Chart 1. *Depositors are strongly urged to meet or exceed these recommendations, especially in the development of prospective digital imaging projects. The Working Group proposes that the Recommended Guidelines become Required Guidelines for projects that begin after 2001.*

Under certain circumstances, digital image materials will also be accepted for deposit that meet only the Minimal Requirements (see Chart 2). It should be understood that such files may place a heavier burden on the central depository staff and these files may need to be reformatted to be managed effectively. *Further, these files may have a relatively short-term life expectancy (e.g., less than 10 years), and may be subject to de-accessioning should the expense associated with maintaining them outweigh the value of preserving and making them accessible.* Only under extraordinary circumstances (e.g., last copy of an item) will digital images files be accepted that do not meet the Minimal Requirements presented in Chart 2.

Table 1: Digital Master Image Files— Recommended Imaging Requirements

Document Type	Resolution	Bit Depth	Enhancements Allowed	File Format	Compression
Printed Text (1)	600 dpi	bitonal	Sharpening, descreening, cropping, deskewing, and despeckling	TIFF 5 & 6	Lossless compression (e.g., ITU-G4)
Rare/damaged printed text	400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Book Illustrations (2)	400 dpi 600 dpi with enhancement	8-gray or 24-color ----- bitonal	Contrast stretching Minimal adjustments for tone and color ----- Descreen/rescreen, sharpen	TIFF 5 & 6	Uncompressed or lossless compression (e.g., ITU-G4, LZW)
Manuscripts	300-500 dpi	8-gray or 24-color, if color present in the original	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Maps & other oversized items	300-400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Graphic Art	400-600 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Photographic Prints	400 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Works of art on paper	400 dpi	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	Uncompressed or lossless compression (e.g., LZW)
Transparencies	4000-5000 on long end or 400 dpi on output > 8" x 10"	8-bit/ channel internal reduction	Contrast stretching Minimal adjustments for tone and color	TIFF 5 & 6	uncompressed or lossless compression; (e.g., LZW)
Microfilm	600 dpi ----- 300-400 dpi at original size	Bitonal ---- 8-bit gray	Sharpening, descreening; cropping deskewing, and despeckling	TIFF 5 & 6	Uncompressed or lossless compression (e.g., ITU-G4, LZW)

Table 2: Digital Master Image Files— Minimal Imaging Requirements

Document	Resolution	Bit Depth	Enhancements Allowed	Format	Compression
Printed Text	300-400 dpi	bitonal	Sharpening, descreening; cropping, deskewing, and despeckling	TIFF 4, 5, & 6 JFIF/ JPEG	lossless or visually lossless compression; e.g., modest JPEG (≥10:1)
	----- 200 dpi	----- 8-gray	----- Minimal adjustments for tone		
Rare/damaged printed text	300 dpi	8-gray or 24-color	Contrast stretching	TIFF 4, 5 & 6 JFIF/ JPEG	lossless or visually lossless compression; e.g., modest JPEG (≥10:1)
	----- 600 dpi	----- bitonal	----- Minimal adjustments for tone and color Sharpening, descreening; cropping deskewing, and despeckling		
Book Illustrations	300 dpi	8-gray or 24-color	Contrast stretching	TIFF 4, 5 & 6 JFIF/ JPEG KPCD	lossless or visually lossless compression; e.g., modest JPEG (≥10:1)
	----- 600 dpi, with enhancement	----- bitonal	----- adjustments for tone and color Sharpening, descreening; cropping deskewing, and despeckling		
Manuscripts	200 dpi	8-gray or 24-color, if color present in original	Contrast stretching	TIFF 4, 5 & 6 JFIF/ JPEG	lossless or visually lossless compression; e.g., modest JPEG (≥10:1)
	---- 400-600 dpi, with enhancement	---- -----1 bit	---- adjustments for tone and color		
Maps & other oversized items	200 dpi	8-gray or 24-color	Contrast stretching, sharpening	TIFF 4, 5 & 6 JFIF/ JPEG	lossless or visually lossless compression; e.g., JPEG (≥10:1)
Graphic Art	300 dpi	8-gray or 24-color	Contrast stretching Documented color correction, sharpening	TIFF 4, 5 & 6 JFIF/ JPEG KPCD	lossless or visually lossless compression; e.g., modest JPEG (≥10:1), Image Pac
Photographic Prints	300 dpi	8-gray or 24-color	Contrast stretching Documented color correction, sharpening	TIFF 4, 5 & 6 JFIF/ JPEG KPCD	lossless or visually lossless compression; e.g., modest JPEG (≥10:1), Image Pac
Works of art on paper	300 dpi	8-gray or 24-color	Contrast stretching, sharpening Documented color correction	TIFF 4-6 JFIF/ JPEG KPCD	lossless or visually lossless compression; e.g., modest JPEG (≥10:1), Image Pac
Microfilm	300-400 dpi -----200 dpi	bitonal -----8-gray	Sharpening, cropping deskewing, despeckling-, Minimal tone adjustment	TIFF JPEG	lossless or visually lossless compression; e.g., JPEG (≥10:1)

C. Quality Control

Quality control (QC) is an integral component of creating digital content that will retain value and utility over time. QC encompasses procedures and techniques to verify the quality, accuracy, and consistency of digital products. The Digital Imaging Tutorial (<http://www.library.cornell.edu/preservation/tutorial/>) outlines the main points of a quality control program. A fully developed strategy for establishing such a program is presented in Oya Y. Rieger, "Establishing a Quality Control Program," in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, pp. 61-83 (Research Libraries Group, 2000). Depositors may request a copy of this chapter from the Department of Preservation by writing to preserve@cornell.edu.

QC Recommendations

1. *Scope of Inspection:* Inspect the quality of the digital image files, the accuracy and consistency of metadata, and the integrity of the storage media for delivery to the central depository.
2. *Extent of Inspection:* Establish a sampling frequency for each aspect of the QC program. Recommended frequency is 100% of all image files and accompanying metadata; minimal requirement is 10% sampling of each image/metadata batch.
3. *Type of Inspection:*
 - a. *Image Files*

The key factors in image quality assessment are resolution, color and tone, and overall appearance. QC can be conducted by visual inspection of images on-screen or via printouts, although it is important to note that quality assessment, especially for tone and color may be highly subjective and changeable according to the viewing environment and the characteristics of monitors and printers. The viewing environment and all links in the imaging system (including the scanner, monitor, and printer) should be carefully controlled. Image quality can also be judged through the use of technical targets (resolution, tone, and color) and increasingly through software. Appendix 1 includes a list of questions to ask in assessing resolution, detail, tone, and color appearance.
 - b. *Metadata*

Metadata has a central role in processing, managing, accessing, and preserving digital image collections. Because of the crucial role it plays in the life cycle of image collections, metadata review should be an integral part of a quality control program. Metadata QC can be automatic or manual or a combination of the two. QC should verify the following: data integrity, form and validity, accuracy of derived data, correctness of data, accuracy and completeness of components, and dynamic metadata. Richard Marisa

describes these aspects and makes recommendations for metadata QC in a sidebar to the chapter referenced above.

4. Pre-Depository Storage and Maintenance Requirements

This section deals with storage and maintenance of digital resources prior to deposit. To ensure long-term viability, transferees shall provide a secure and reliable storage environment for the digital files. Good storage practice plays a key role in preventive preservation, which is a crucial strategy to control and reduce resource requirements associated with preservation. The following list highlights the key requirements for proper storage and maintenance of digital collections that may ultimately be transferred to the central depository.⁷

1. *Media:*

- Store master files on high quality, industry standard digital tape, magnetic disks, CD-R, or other contemporary media approved by the Depository staff.
- Check media periodically for readability depending on the manufacturer's recommendations.

2. *Backups:*

- Create backups of the master files and store off-site in a secure location.

3. *Recording/Reading Devices:*

- Monitor the recording and access devices, such as tape drives, and make sure that they are of good quality and well-maintained (note that problems with the access devices e.g., head/media crashes are one of the most common causes of damage to magnetic storage media).

4. *Storage:*

- Store media in a controlled environment. The accepted ranges for temperature are 62°-68° (65° optimum) and for humidity are 35%-45% (40% optimum).
- Establish consistent levels within the acceptable range (this is more important than attempting to maintain the optimum temperature and humidity.)
- Store media away from strong magnetic fields.
- Maintain a clean operating environment.
- Minimize the handling and use of magnetic storage media to reduce wear.
- If media is stored off-line, store it vertically in appropriate containers.

⁷ The resources used in developing this section included:

Maggie Jones and Neil Beagrie, *Preservation Management of Digital Materials Workbook*, 2000, www.jisc.ac.uk/dner/preservation/workbook/

Neil Beagrie and Daniel Greenstein, *A Strategic Policy Framework for Creating and Preserving Digital Collections*, Arts and Humanities Data Service Executive, 1998, ahds.ac.uk/manage/framework.htm.

J. Van Bogart, *Magnetic Tape Storage and Handling*, CLIR, 1995, www.clir.org/pubs/reports/pub54.html

5. Security:
 - Control access to storage facilities and processing areas. Store media in a separate, preferably lockable area.
 - Employ appropriate security systems and procedures to protect the authenticity of the collections and ensure no deliberate or inadvertent changes take place.
6. Refreshing:
 - Digital files maintained for an extended period prior to deposit should be refreshed to new media regularly, taking into consideration the recommendations of the media supplier for certain environmental conditions and following the trends for more efficient storage technologies (e.g., refreshing may be necessary when new storage systems are purchased).
 - Follow a verification procedure such as checksum or MD5 to ensure the authenticity and integrity of the files after media refreshing.
7. Documentation:
 - *Document actions taken during refreshing or other maintenance operations that may affect the integrity of files.*

5. Metadata Requirements

A. Descriptive Metadata

Descriptive metadata, loosely defined as information used by the delivery system for resource discovery and to identify resources, are an integral part of any digital access system and are essential to the long-term maintenance of digital files. The Cornell University Library currently uses the online catalog as its database of record for descriptive metadata. This will continue to be the case as the library begins its work of providing long-term access to digital files. However, depending on the nature of the digital material, additional descriptive metadata records may be needed to support fully this endeavor.

Title-level MARC records should be created for all monographs and serials that are added to the digital depository. As has been observed repeatedly with aggregations of electronic journals and e-books, the presence of title-level MARC records in the online catalog will increase the use of these resources. Furthermore, records for these materials will be shared in the national utilities and may be used by other institutions to provide access to resources housed at Cornell. In addition to increasing use, MARC records can be used to form the basis of descriptive metadata used in an electronic delivery system - either through links to a system or through the duplication of a portion of the MARC data in the digital system. Although particular procedures and standards have changed over time, all MARC records created should follow current cataloging standards and should meet at least the minimal requirements defined by current, local procedures. Appendix 2 includes an example of descriptive metadata. Cornell's policies for cataloging digital materials are embodied in:

Cataloging Procedures for Networked Electronic Resources
<http://www.library.cornell.edu/voyager/Bibs/ECat/e-catTOC.html>

These procedures provide guidelines for handling basic issues as well as using separate and multiple-version records, treating multiple electronic versions, and other complex situations. As noted in section 2.1 of the *Cataloging Procedures for Networked Electronic Resources*:

This document provides local usage guidelines for cataloging networked electronic resources in Voyager and in the Library Gateway. It does not cover everything one needs to know to process these items. For a complete list of field definitions, appropriate tags, and national standards, consult USMARC and CONSER documentation. For more general instructions on the cataloging of networked e-resources, see Nancy Olson's *Cataloging Internet Resources: A Manual and Practical Guide*, 2nd ed.

This advice should also be followed when approaching the descriptive cataloging of digital image materials. Although local practice takes precedence, catalogers may also consider the following publication:

Library of Congress, *Draft Interim Guidelines for Cataloging Electronic Resources*, http://lcweb.loc.gov/catdir/cpso/elec_res.html

In addition to following the standards listed in these procedures, cataloging staff should be certain to take the following into account:

- Particular attention should be paid to the use of the 899 "Local series code" field. A unique code is to be created for each digital collection and recorded in the 899 field of each record that reflects a title in that collection. This code will collate each collection of MARC records in the online catalog to assist in record maintenance, data migration, and other functions. Instructions for assigning 899 codes are available in *Cataloging Procedures for Networked Electronic Resources*. Establishing an 899 code should be done in cooperation with staff from the Technical Services Support Unit.
- Records should use appropriate 007 values according to MARC 21 <http://lcweb.loc.gov/marc/bibliographic/ecbd007s.html>. In 2000, many 007 values were enhanced to capture needed data concerning preservation and reformatting issues and to record image production values. Adding this information to previously cataloged items may not be practical but is desirable if possible. In addition, all prospective cataloging should take advantage of this useful field.
- Special attention should be given to LCRI 1.11A "Facsimiles, photocopies, and other reproductions" (revised Summer, 2000). This LCRI provides assistance in handling non-microfilm reproductions of print materials (including digital reproductions) and suggests methods for recording information concerning the digitization process. Cataloging staff should also follow current, local procedures for handling such

reproductions. In particular, be sure to include either a 533 "Reproduction Note" in the bibliographic record or an 843 "Reproduction Note" in the holdings record. This information is essential in tracking the agency responsible for the creation of the reproduction.

- Notes on image file format are routinely added to the MARC record. This information is likely also to be contained in administrative metadata that will be part of a digital collection management system. When adding this information, be aware that if materials are migrated, this information will become outdated and may need to be changed in the MARC record.
- Cornell University Library has decided that persistent, stable URLs should be used in most MARC records for electronic resources, such as digital image files, and has chosen OCLC's PURL server as the method for creating such persistent identifiers. URLs that appear in the record may need to have PURLs created for them. The procedures for creating PURLs (including information on resources that do not require PURLs) are defined in Section 6 of the *Cataloging Procedures for Networked Electronic Resources*:
<http://www.library.cornell.edu/voyager/Bibs/ECat/e-cat6.html>.
- The Technical Services Support Unit handles the creation and maintenance of PURLs and should be consulted about setting up any needed PURLs. PURLs for collections of digital resources utilize the local series code listed in the 899 as well as the online catalog bibliographic record ID. The creation of PURLs for these materials is important. The PURL will reside in the online catalog and will also be part of all records added to the bibliographic utilities. This is particularly important for long-term access since image file locations are likely to change over long periods of time.

As cataloging issues arise, these documents will change and other documents will emerge to standardize the treatment of electronic resources. If you are unfamiliar with the treatment of these materials, be sure to check with other local catalogers about the most current methods for approaching the creation of MARC records for this type of material.

Unlike monographs and serials, many materials (such as archival collections, pamphlets, manuscripts, letters, photographs and other visual resources) often do not have title-level MARC records created for them. In many cases, there is a one-to-many relationship between the MARC record and the material itself. These materials present a unique challenge to the long-term accessibility of descriptive metadata. Although title-level records must be created for monographs and serials, they are not mandatory for other materials but do represent an ideal level of access. These materials must have at least collection-level MARC records. These records should follow the standards outlined above as closely as possible given the differences inherent in the material. More detailed descriptive metadata will be needed for the creation of adequate access systems for these materials. The record structure of this metadata is likely to be unique and idiosyncratic. The same is true of article-level (or chapter-level) metadata that might be recorded for electronic serial (or monograph) collections. This situation will present difficulties for the long-term maintenance of this data. As soon as it is practical and possible, idiosyncratic data structures should be migrated to more universal standards. For instance, the use of Dublin Core may be preferred as a long-term storage structure for

bibliographic data over a homegrown structure. However, this will only be practical when access systems are created to utilize metadata stored and structured in such a way.

When archival materials are digitized, that process should follow the creation of a collection finding aid, or the recording of equivalent information, and should be guided by that finding aid, which will contain the most comprehensive descriptive metadata about the collection. Archival finding aids should be expressible with the Encoded Archival Description DTD. Although the finding aid will contain information about the collection's physical organization or location, it should ideally describe the intellectual arrangement of the collection. The finding aid should be viewed as the principal access and navigation mechanism into the images. Collection-level MARC records should link to the finding aid when possible. To help track the relationship among the digital images, the EAD finding aid or equivalent, and the collection-level MARC record, the EAD finding aid or equivalent should include the MARC record bibliographic record ID and/or the collection number. Since library management systems will change over time, the collection number is likely to remain a stable identifier.

The Rare and Manuscript Collections has begun to define an "RMC EAD template," which can be found at <http://cidc.library.cornell.edu/xml/template/>. As this document evolves, it will define minimum field/element requirements, as well as a set of local content standards. In addition, it should reflect wider archival community standards concerning the interoperable use of EAD.

Following the criteria described above is necessary not simply for information retrieval and access but also for the maintenance of descriptive metadata. Since MARC is a rich, descriptive metadata standard, it is possible to crosswalk portions of this data to less-complex, short-term formats used by access systems. Unlike metadata structures like the Dublin Core, MARC has a clear process for supporting and maintaining both the form and content of records, can be shared easily among institutions, and is widely recognized, supported, and documented. Non-MARC descriptive metadata used in a delivery system should make reference to the MARC record(s) associated with the material, preferably by citing the online catalog record ID(s). This will enable CUL to update delivery system metadata to capture changes in names, titles, and subjects as well as new standards, such as Unicode. In addition, as common, non-MARC metadata standards develop, MARC will be used to populate records that follow those standards when possible to limit the amount of customized data migration that will need to take place.

If CUL widely adopts a metadata structure that can be used in place of MARC, materials submitted to the depository may need records in that structure as well as or instead of MARC records. In addition, as metadata formats and standards emerge, CUL may migrate all MARC data into other metadata structures for long-term preservation of descriptive metadata.

B. Structural Metadata

Structural metadata provides essential information needed to guarantee adequate long-term access to digital image files. Although less important for single-image objects (such as photographs, artwork, posters, or maps), structural metadata is invaluable for access to objects comprised of more than one image or for single images that are related to other objects (such as a series of letters or memos meant to be read in tandem). Unfortunately the way that structural metadata is stored will be dependent on the system used to provide access to the image material. However, each collection's structural metadata should be captured and stored in a uniform manner. This is needed to create stable and consistent access systems as well as to preserve both metadata and data files.

As part of the IMLS grant "Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy" (1999-2000), the project team defined a series of structural elements that were identified as mandatory (M), mandatory if applicable (MA), and optional (O) for digital image collections at Cornell. Table 3 was generated by examining various preservation metadata proposals for descriptive and administrative metadata to facilitate preservation decisions.

Table 3: List of Structural Metadata Elements for Digital Image Collections

M = Mandatory; MA = Mandatory if applicable; O = Optional

All Materials

Relationship to Other Resources (MA)
Metadata Locations (M)
Start image/page (M)
End image/page (M)

Manuscripts

Title page (MA)
Colophon (O)
Caption (O)
Heading (O)
Leaves (O)
Illustrative matter (O)
List of illustrations (O)
Table of contents (MA)
List of tables (O)
Page numbers (M)
Blank page (M)
Marginalia (O)
Front matter (MA)
Back matter (MA)

Monographs

Title page (M)
Copyright page (M)
Table of contents (M)
List of illustrations (O)
List of tables (O)
Beginning segments (e.g., forward, preface, acknowledgements) (O)
End segments (e.g., epilogue, afterword, conclusion, etc.) (O)
Chapters/parts (O)
Notes (O)
Bibliography (O)
Index (M)
Colophon (O)
Errata (O)
Page numbers (M)
Blank page (M)

Pamphlets

Title page or cover (M)
Copyright page (MA)
Table of contents (MA)
List of illustrations (O)
List of tables (O)
Beginning segments (e.g., forward, preface, acknowledgements) (O)
End segments (e.g., epilogue, afterword, conclusion, etc.) (O)
Chapters/parts (O)
Notes (O)

- Bibliography (O)
- Index (MA)
- Colophon (O)
- Page numbers (M)
- Blank page (M)

Serials

Entire Publication

- Volume (M)
- Issue (M)
- Supplements (M)
- Table of contents (M)
- Index (at issue and volume level) (M)
- Corrections and retractions (O)
- Serial front matter (M)
- Serial part (O)
- Serial section (O)
- Name index (if separate from other index) (O)
- Subject index (if separate from other index) (O)
- Errata (O)
- Page numbers (M)
- Blank page (M)

Articles

- Article title (O)
- Author (O)
- Abstract (O)
- Date (O)
- Tables/figures (O)
- Errata (O)
- Page numbers (M)
- Blank page (M)

C. Preservation Metadata

Preservation metadata encompasses a range of information that is required for the short- and long-term management of digital image files. This category of metadata includes both micro information that describes the technical specifications of an image collection as well as administrative information that will support future preservation decision making and action. This section outlines the preservation metadata requirements in two categories. The first section, Digital Image Collections Inventory, outlines an information system that aims to collect and maintain high-level administrative metadata. The second category, Technical Metadata, attempts to collect technical information at a micro-level.

Digital Image Collections Inventory

The goal of the Digital Image Collections Inventory database is to provide profiling information on CUL's digital image collections. It aims to include general information that will support preservation administration and decision-making. This inventory approach tries to address the difficulty associated with gathering base-level information about the library's individual image collections. Every project team is required to complete this questionnaire during the project implementation phase as the collection of elementary information has proven to be very difficult to recreate after the fact. Table 4 lists the recommended data elements for such an inventory database. Currently, this type of information is not required of staff members who are involved in different stages of digital imaging projects. There is no formal recording or sharing obligation. Although this questionnaire does not address the need for detailed administrative metadata (and how it is recorded and maintained), it presents an easily attainable and effective approach for short-term management of administrative metadata until we have more sophisticated systems and standards in place. The technical metadata standard that is currently being developed by ANSI/NISO (<http://www.niso.org/commitau.html>) will complement this approach by providing a framework to record in-depth information on the technical specification of individual collections.

The short-term plan to implement this inventory is through the development of a Web survey (with an automated recording system). However, the long-term recommendation for the collection and manipulation of such data is to create a DTD for XML (or an XML schema) implementation. Project coordinators and participants can easily collect the information elements requested in the inventory, with an estimated time involvement of one hour. However, some of the required information is dynamic and cumulative (e.g., refreshing and migration history) and therefore would require ongoing updates. After the completion of the key sections by project staff during the implementation phase, the central depository staff may need to update different parts of the inventory throughout the life cycle of a particular collection. The suggested frequency for the revision of the dynamic fields is one year.

Table 4: Digital Image Collections Inventory: Data Fields

<i>Project description</i>
project title
year the collection was created
project leaders/coordinators, team members
project partners
sources of funding
reason for the project
<i>Source type and characteristics</i>
document type (e.g., printed text, book illustrations, color photographic prints, manuscripts, etc.)
physical dimensions (category: regular, oversize - if possible exact size, or "size varies" statement with min and max measurements -, size varies, but no greater than 8.5 x 11 or some such)
scanned from original or film intermediate
subject matter
<i>Collection size</i>
total file size of the collection including image and metadata files, programs, scripts, etc. (estimated or actual)
number of images
<i>Storage media</i>
type and location
<i>Scanning information</i>
resolution
bit depth
color space or CLUT information for color documents
file format and version
compression technique, version, and ratio
scanner used
vendor vs. in-house scanning
<i>Processing information</i>
any image enhancements on the master copy? E.g., how were halftones handled? Any special treatment?
derivatives created (access, processing; such as scaled/reformatted copies for Web delivery, OCR'ed images, etc.)
<i>Metadata</i>
file header (if possible tags used)
what kind of descriptive metadata – where and how recorded? (e.g., MARC, Dublin, PURL, etc.)
what kind of structural metadata – where and how recorded? (SGML, XML, structuring tags, external metadata, etc.)
what kind of technical metadata – where and how recorded?
special collections – finding aid information
<i>Access mechanisms</i>
online/offline
Web address

<i>System/interface design and characteristics</i>
system specifications (e.g., based on Hunter, OpenText, etc.)
known system requirements
key interface features (forms and style sheets, use of JavaScripts, etc.)
<i>Refreshing/migration history</i>
<i>Rights management & Authenticity</i>
document the process of clearing copyright issues
license information
display and transmission restrictions, right holders
any security/authenticity measures (e.g., watermark)
chain of custody

Appendix 3 includes a sample entry for the Save America's Treasures collection to demonstrate the use of this inventory.

Technical Metadata

One of the key requirements of a preservation policy is to have a framework for collecting and recording technical metadata to safeguard information that may be essential in monitoring and rescuing files in the face of changing technologies. Continued viability of the CUL digital image collections heavily depends on the availability of information on technical characteristics of collections, technological dependencies, change history, and rights management. Technical metadata serves several purposes. In a managerial context, it supports image quality assessment, image enhancement and processing, and facilitates work-flow management. Although there is limited evidence at this point, technical metadata is also seen as an important source for long-term collection management. The Cornell University Library intends to adopt the technical metadata standard that is currently being developed by NISO. Standardization of technical metadata will facilitate a systematic approach in recording and managing technical image information. The charge of the Technical Metadata for Digital Still Images Standards Committee (<http://www.niso.org/commitau.html>) is to review and revise the Data Dictionary for Technical Metadata for Digital Still Images (Working Draft, 1.0, July 2000). The proposed data dictionary presents a comprehensive list of elements required to describe the technical features of image files. The data fields are organized in four groups: basic image parameters, image creation, image performance assessment, and change history. The standard in development is not fully addressing the implementation question so it will become a CUL-based decision whether to develop and adopt a DTD for technical metadata to support an XML implementation.

Authenticity

Master digital images deposited in the Cornell Digital Library should represent accurate, complete, and trusted versions of the original source materials. Second, they should have documentation demonstrating an unbroken chain of custody since their creation. Third, master digital files deposited in the Cornell Digital Library should have documentation demonstrating how the files have been protected from un-documented or unintentional change, such as tampering. This documentation should cover the technical procedures followed to ensure that the files retain quality, integrity, authenticity, and reliability after creation. If more than one version of the digital image materials exists, the version that has been most closely monitored and safeguarded will be the preferred version for deposit. Once deposited, the Depository staff shall continue to document the chain of custody and assume responsibility for safeguarding the authenticity of the digital collections.

ROLE AND RESPONSIBILITIES OF A CENTRAL DEPOSITORY

This report began with a strong recommendation to the Library Management Team to establish centralized responsibility for ensuring continuing access to digital image collections over time. This responsibility should take the form of a Central Depository that is administratively located within the Digital Library and Information Technology (D-LIT) infrastructure. The Central Depository's role will be to facilitate the long-term management and use of digital resources in the most cost-effective manner, based on the distinctive characteristics of the Cornell University Library system.

A detailed description of the responsibilities of the Central Depository will constitute the second part of this document, and will be prepared after the proposed feasibility study (see below). This forthcoming section will focus on the responsibilities of the Central Depository, and will cover the following key issues:

- Guidelines on transmission methods, documentation, and media for deposit
- Acquisition procedures and protocols to: verify the arrival, completeness, validation, and readability of deposited material; reformat or copy materials to new media; and provide for inventory control
- Maintenance/updating of depository guidelines
- Respective responsibilities and rights of the transferee and the depository staff, including on-going interactions
- Outreach to potential transferees
- Relationship to Digital Library staff (access and availability, derivative creation), collection development (selection), Technical Services (documentation, cataloging, and creation of PURLs), Access Services (technical reference and user support), and Preservation (policy development)
- On-going system maintenance practices, including
 - Storage, backup, and redundancy procedures
 - Data security, integrity, and auditing requirements
 - Collection monitoring
 - Media refreshing
 - Disaster recovery plans
- Upgrading/modification procedures (e.g., replacement images, metadata updates, alternative access versions)
- Rights management (including authorized use of software and content for preservation purposes, creation of rights clearance forms)
- Outsourcing, contractual arrangements, and collaboration with other units/institutions (offsite storage, redundancy requirements)
- Access policies (in conjunction with Digital Library staff)
- Preservation alternatives, including risk assessment associated with various strategies
- Technology monitoring
- Documentation, including maintenance of the image collections inventory
- Collection review; deaccession and disposition guidelines
- Resource requirements, including cost assessments

NEXT STEPS

A. *Needs Assessment Survey*

The guidelines presented in this document represent an important first step in developing an institutional digital preservation strategy. A necessary next step is to address the managerial processes of the depository, including the day-to-day and long-term technical and financial management of data. To be able to accurately describe the role of the central depository and its processes, the Working Group recommends that the library conduct a preservation readiness feasibility study. Such a study will be instrumental in identifying the central depository requirements for image collections and also understanding the CUL's existing technical and administrative infrastructure necessary for the attainment of these objectives. The goal of the feasibility study is to explore the following issues:

- 1) Explore the short- and long-term role of a central depository for digital collections
 - Identify the role and working relationship of the key players
 - Evaluate the existing organizational structure and staffing patterns to assess the readiness of CUL to fulfill preservation requirements.
 - Identify resource requirements including staffing (level and skills), equipment and space needs, financial planning, CUL-CIT collaboration, etc.⁸
- 2) Evaluate the relevance and applicability of various institutional policy frameworks, preservation architectures and procedures (e.g., assessment of OAIS and Encompass for long-term maintenance purposes) for CUL
- 3) Assess the effectiveness of current maintenance practices for long-term preservation
- 4) Evaluate the extensibility of the policies developed for digital image collections in meeting the needs of CUL's other digital collections and initiatives (e.g., Euclid, Harvest, and Prism)
 - Identify the divergence and convergence of preservation policies and frameworks for different digital formats (e.g., images, HTML files, numeric files, etc.)
- 5) Consider the use of a rating system to identify the permanence level of CUL's electronic resources (e.g., see the National Library of Medicine's permanence rating model for electronic resources, www.arl.org/newsltr/212/nlm.html)

⁸ After a thorough analysis of the existing cost studies for digital preservation, the group decided that there were no existing models that could be readily implemented for our purposes. The main challenge is that most of the existing studies are not itemized and do not indicate what is included in the cost estimates. In addition, most of them present preservation costs on a per gigabyte basis, focusing mostly on storage costs. There are considerable economies of scale in large archives, so calculating costs based on a gigabyte unit may not be accurate as a system continues to grow. Expenses for digital preservation start accumulating soon after selection for digitization and continue as long as access to the digital collection must be ensured.

B. Schedule and Procedures for Updating the Deposit Guidelines

The members of the Digital Preservation Policy Working Group will reconvene once a year to review the document and to identify sections that need to be updated. The co-chairs of the committee will continue to organize the updating process by convening these annual meetings and also following up after the meetings to reflect the recommendations to this document.

During the first year of implementation of the depository, the Digital Preservation Policy Working Group will continue in an advisory role in the development of the depository guidelines.

C. Publicity and Training

After this document is approved by the Library Management Team, it will be distributed to the library staff for comments and questions. In addition, there will be a number of orientation and training sessions to familiarize the library staff with the requirements articulated in this document. The Digital Imaging and Preservation Unit of the Department of Preservation and Conservation Department offers several educational materials and can accommodate Cornell staff in their week-long digital imaging workshops (<http://www.library.cornell.edu/preservation/workshop/>). The following online tutorial prepared by the department will be useful in introducing the staff to basic concepts related to creating and managing digital image collections:

Moving Theory into Practice: Digital Imaging Tutorial

<http://www.library.cornell.edu/preservation/tutorial/>

Appendix 1: Image Quality Assessment

This section is reproduced from:

Oya Y. Rieger, "Establishing a Quality Control Program," in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, pp. 61-83 (Research Libraries Group, 2000).

A. Questions to Ask in Evaluating Resolution and Detail

Compare the digital images (or their printouts) to the original documents (or to the intermediates):

Text/line Art Documents

1. Is the stroke adequately reproduced?
2. Is the significant detail adequately reproduced?
3. Is the smallest text readable?
4. Are individual line widths (thick, medium, and thin) rendered faithfully?
5. Are serifs and fine detail rendered faithfully?
6. Are adjacent letters as separate as they should be?
7. Are the open regions of lowercase characters retained (i.e., not filled in)?
8. Are the edges of individual letters or shapes as smooth or well defined (not ragged) as the original?⁹
9. Is there good contrast or differentiation between the text and the background?
10. Is there even illumination across the image (i.e., is the image washed out or too dark)?
11. Is there a gray cast or streaking in the background?
12. Is the document fully reproduced?

Continuous-Tone and Halftone Documents

(The first three questions apply only to continuous-tone documents.)

1. Is the stroke adequately reproduced?
2. Is the significant detail adequately reproduced?
3. Is fine detail in the darkest and lightest portions retained?
4. Are there even gradations across the image (e.g., no banding, streaking, newton rings, or graininess)?¹⁰

⁹ Edge raggedness relates to the smoothness or straightness of edges along lines at very close inspection. Pay special attention to curved and diagonal lines on characters and line graphics.

¹⁰ Streaking and graininess are typical film attributes that might be evident in the displayed image. Banding (varying lightness and darkness) may be attributable to improper lighting. Newton rings (circular impressions) can be introduced during the scanning of transparencies.

5. Is the image free of a moiré effect?¹¹
6. Is the significant informational content adequately reproduced?
7. Is the document fully reproduced?
8. Is the image too light or too dark?

B. Questions to Ask in Evaluating Color and Tone Appearance

Compare the image to the original document, an intermediate, or color/grayscale bars.

Grayscale Images

1. Evaluate tone appearance in the highlights (lighter sections), midtones, and shadows (darker sections). Are the details in these different sections captured without any loss?
2. Is the image too light or dark overall?

These questions are based on the grayscale targets used in photography and scanning:

3. How many grayscale bars can you count on your grayscale image?
4. If your grayscale target is numbered, at what numbers do you cease to discern distinct shades of white, gray, and black?
5. Is there an overall color shift on the grayscale target?
6. Use the information option of your image viewing software to read the color (RGB) values presented at different grayscale steps. How do they compare to the reference values provided by the grayscale bar? What is the difference between the smallest and largest values for each color channel for individual color patches (variance indicates color imbalance)?
7. Display a histogram of your grayscale bar image.¹² Are all digital levels from 0 to 255 used? Do you observe any clipping?

Color Images

1. Do you observe a color shift in the overall image or an obvious shift to a certain color?
2. Study the red, green, blue, and yellow colors. Do any show a color shift? Is it minimal or obvious?
3. Evaluate colors in the highlights, midtones, and shadows, especially red, green, blue, and yellow. Do any show a color shift? Is it minimal or obvious?
4. Is the image light or dark overall?

¹¹ Moiré patterns are most noticeable in the lighter regions of an image and in areas of “low activity” (e.g., in the sky portion of a landscape halftone). In portions with busy content (high activity), moiré is often hidden. Evaluate halftones onscreen only at 1:1 (100%); any other view might introduce halftone patterns not native to the image file.

¹² Rely on your image viewing software's user guide to find out how to create and evaluate histograms.

These questions are based on the grayscale and color targets used in photography and scanning:

5. How many grayscale bars can you count on your grayscale image?
6. If your grayscale target is numbered, at what numbers do you cease to discern distinct shades of white, gray, and black?
7. Do you notice an overall color shift on the grayscale or color target? If so, does it fall within your tolerance range?
8. As shown in figure 4, use the Window/Show Info option of the Adobe Photoshop software to read the color (RGB) values presented at different grayscale steps. How do they compare to the reference values on the grayscale bar? What is the difference between the smallest and largest values for each color channel for individual color patches?
9. Use the hue, saturation, and brightness adjustments of your image viewing software to evaluate the individual colors of the color bar. Comparing the colors on the color target to the original color target, is there a color shift to a certain color? Is it minimal or obvious?
10. Even if the color bar evaluation is satisfactory, compare different sections of the document to the image: is the color satisfactory?

C. Overall Evaluation

The final overall evaluation of an image should combine all the individual factors that contribute to its quality, such as capture system performance, resolution, dynamic range, and color accuracy. Your subjective evaluation should confirm your objective conclusion that an image is satisfactory. Remember that it is *impossible* to generate an image that will fully replicate the look and feel of a document. Ask these questions to evaluate overall image quality:

Does the image convey all the significant information included in the original document (e.g., translucency in a watercolor painting, overlay in an oil painting, quality and texture of paper, etc.)? If not, how much does this affect your satisfaction with the image?

Compared to the original document, is the image:

- unacceptable?
- adequate but diminished?
- comparable?
- improved?

Will the user be satisfied with the image as a document surrogate, or will the image serve just as a basic access tool?

Even if the image passes your subjective and objective inspection based on the grayscale and color targets:

- Is the image's overall dynamic range adequate?
- Are you satisfied with the general color appearance of the image?

Appendix 2: Descriptive Metadata Example

MOA Multiple-Version OPAC View

Author/Creator: United States. Naval War Records Office.

Title: Official records of the Union and Confederate Navies in the War of the Rebellion.

Published: Washington, Govt. Print. Off.,

Description: 30 v. : ill., maps (part fold.) ports. ; 23 cm. index.

Electronic Access: <http://resolver.library.cornell.edu/moap/anu4547>

Subjects:

United States. Navy--History--Civil War, 1861-1865.

Confederate States of America. Navy--History.

United States--History--Civil War, 1861-1865--Naval operations--Confederate States.

Other Names: United States. Office of Naval Records and Library.

Series:

Office memoranda (United States. Naval War Records Office)

Office memoranda (United States. Naval War Records Office).

Notes:

Issued in the congressional series as House documents.

Series 1, v.1-27; Series 2, v.1-3.

Indexes: Ser. 1, v.1-13. 1 v. (Issued as the Office's Office Memoranda) (E591.U575)

Ser. 1, v.1-ser. 2, v.3. 1 v.

Location: *Networked Resource

Call Number: ONLINE

Status: Available

Volumes : Ser.1-2

Indexes: Index

Reproduction Note:

Computer file. Ser.1-2. Ithaca, N.Y. : Cornell University Library, 1995. [27,515] image files.

Notes:

Files for the images of individual pages are encoded in Aldus/Microsoft TIFF Version 5.0 using facsimile-compatible CCITT Group 4 compression.

Location: Olin Library

Call Number: E591 .U58

Copy Number: 2

Status: Available

Indexes: v.1

Location: Olin Library

Call Number: E591 .U58 1894a

Status: Available

Volumes : Ser.1-2

Indexes: Index

Reproduction Note: Reproduction from digital master. Ser. 1-2. Ithaca, N.Y. : Cornell University Library, 1995. 23 cm.

MOA multiple-version MARC record

000 01515cam 2200337 a 450

001 2797023

005 19990908120000.0

008 870716m18941922dcuabc f001 0 eng d

010 __ |a 06035188
 035 __ |a (NIC)notisANU4547
 040 __ |a NIC |c NIC
 043 __ |a n-us--
 110 1_ |a United States. |b Naval War Records Office.
 245 10 |a Official records of the Union and Confederate Navies in the War of the Rebellion.
 260 __ |a Washington, |b Govt. Print. Off., |d 1894-1922.
 300 __ |a 30 v. : |b ill., maps (part fold.) ports. ; |c 23 cm. |e index.
 490 1_ |a U. S. Naval War Records Office. Office memoranda
 500 __ |a Issued in the congressional series as House documents.
 500 __ |a Series 1, v.1-27; Series 2, v.1-3.
 500 __ |a Indexes: Ser. 1, v.1-13. 1 v. (Issued as the Office's Office Memoranda) (E591.U575)
 500 __ |a Ser. 1, v.1-ser. 2, v.3. 1 v.
 610 10 |a United States. |b Navy |x History |y Civil War, 1861-1865.
 610 10 |a Confederate States of America. |b Navy |x History.
 651 _0 |a United States |x History |y Civil War, 1861-1865 |x Naval operations |z Confederate States.
 710 1_ |a United States. |b Office of Naval Records and Library.
 830 _0 |a Office memoranda (United States. Naval War Records Office)
 830 _0 |a Office memoranda (United States. Naval War Records Office).
 899 _0 |a MOAProj
 856 40 |u <http://resolver.library.cornell.edu/moap/anu4547> |x <http://moa.cit.cornell.edu/MOA/MOA-JOURNALS2/OFRE.html>

Appendix 3: Sample Entry for Digital Image Collections Inventory

Save America's Treasures

Cornell University Library, Preservation and Conservation Department
[12/18/00 - Robert S. Glase]

Project description

<u>Project title:</u>
<u>Save America's Treasures</u>
<u>Year the collection was created:</u>
<u>September 2000</u>
<u>Project leaders/coordinators, team members:</u>
<u>Barbara Berger Eden, John Dean, Elaine Engst, Preservation-Conservation staff</u>
<u>Project partners:</u>
<u>CUL Preservation-Conservation, CUL Division of Rare and Manuscripts Collections</u>
<u>Sources of funding:</u>
<u>Save America's Treasures Grant; Total: \$662,000</u> <u>Direct Funds: \$331,000; Cost share \$331,000</u>
<u>Reason for the project:</u>
<u>Digitization of a historically significant anti-slavery collection</u>

Source type and characteristics

<u>Document type (e.g., black and white text-based material, heavily illustrated text, color photographs, special collections materials, slides, etc.):</u>
<u>Black and White, some illustrations/engravings</u>
<u>Physical dimensions (category: regular, oversize - if possible exact size, or "size varies" statement with min and max measurements -, size varies, but no greater than 8.5 x 11 or some such):</u>
<u>Average 4 x 7</u>
<u>Scanned from original or film intermediate:</u>
<u>Original</u>

<u>Subject matter:</u>
<i>Abolition of slavery in United States, Conditions of slaves in United States (early nineteenth century)</i>

Collection size

<u>Total file size of the collection (estimated or actual):</u>
<i>Project is approximately 20% complete: 4861.0 MB</i>
<u>Number of images:</u>
<i>Project is approximately 20% complete: 42,032</i>
<i>The total has been estimated at around 700,000 images.</i>

Storage media

<u>Type and location:</u>
<i>None - direct to library server</i>

Scanning information

<u>Resolution:</u>
<i>600 dpi bitonal</i>
<i>400 dpi grayscale</i>
<u>Bit depth:</u>
<i>600 dpi bitonal-1 bit</i>
<i>400 dpi grayscale-8 bit</i>
<u>Color space or CLUT information for color documents:</u>
<i>N/A</i>
<u>File format and version:</u>
<i>Tiff 6.0</i>
<u>Compression technique, version, and ratio:</u>
<i>N/A</i>
<u>Scanner used:</u>
<i>Xerox DocuImage 620s</i>
<u>Vendor vs. in-house scanning:</u>
<i>In-house</i>

Processing information

<u>Any image enhancements on the master copy? (E.g., how were halftones handled? Any special treatment?):</u>
<i>400 dpi grayscale used as needed for halftones</i>
<u>Derivatives created (access, processing; such as scaled/reformatted copies for Web delivery, OCR'ed images, etc.):</u>
<i>None to date (note: to be created by RMC)</i>

Metadata

<u>File header (if possible tags used):</u>
<i>N/A</i>
<u>What kind of descriptive metadata – where and how recorded? (e.g., MARC, Dublin, PURL, etc.):</u>
<i>MARC-To be created by RMC</i>
<u>What kind of structural metadata – where and how recorded? (SGML, XML, structuring tags, external metadata, etc.):</u>
<i>Image tags to be created</i>
<u>What kind of technical metadata (where and how recorded?):</u>
<i>File size, image count. Recorded in Filemaker Pro.</i>
<u>Special collections – finding aid information:</u>
<i>N/A</i>

Access mechanisms

<u>Online/offline:</u>
<i>Online</i>
<u>Web address:</u>
<i>None to date</i>

System/interface design and characteristics

<u>System specifications (e.g., based on Hunter, OpenText, etc.):</u>
<i>MOA specifications (note: to be implemented by David Ruddy.)</i>
<u>Known system requirements:</u>
<i>Web interface through CUL digital library</i>
<u>Key interface features (forms and style sheets, use of JavaScripts, etc.):</u>

<u>None</u>

Refreshing/migration history

<u>Refresh</u>
<u>None to date</u>
<u>Migrate</u>
<u>None to date</u>

Rights management & Authenticity

<u>Document the process of clearing copyright issues:</u>
<u>To be resolved by RMC</u>
<u>License information:</u>
<u>None</u>
<u>Display and transmission restrictions, right holders:</u>
<u>None</u>
<u>Any security/authenticity measures (e.g., watermark):</u>
<u>None</u>
<u>Chain of custody:</u>
<u>RMC</u>