Digitization and Digital Preservation

By Steven Puglia
Preservation and Imaging Specialist
U.S. National Archives and Records Administration
8601 Adelphi Road, College Park, MD 20740, USA
Phone: 301-837-3616
Email: steven.puglia@nara.gov
April 2010

Digitization is more than scanning.

- HANDBOOK FOR DIGITAL PROJECTS: A Management Tool for Preservation and Access
- Moving Theory into Practice - Cornell
- JISC Digital Media
- Cataloging & Digitizing Toolbox - LC

Digitization is more than scanning-

A complete process that broadly includes:

- Selection
- Assessment
- Prioritization
- Project management and tracking
- Preparation of originals for digitizing

- Metadata collection and creation
- Digitizing
- Quality management
- Data collection and management
- Submission of digital resources to delivery systems and into a repository environment
- Assessment and evaluation of the digitization effort

NISO - A Framework of Guidance for Building Good Digital Collections

3rd Edition, 2007

Digitizing Equipment

Essential Characteristics -

- What are the essential characteristics of the originals that we want to replicate and carry forward?
- What characteristics can we stand to lose?
- How much information needs to be captured?

Need to define essential characteristics of the original resources -

- Microfilm standards and guidelines focus on maintaining text legibility.
- Specifications for photographic duplicates define approaches to produce duplicates that have the same photographic properties as the originals - same overall density, density range, and relationship between the tones.

Essential Characteristics -

- Characteristics informed by functional, technical, physical, qualitative, curatorial, archival, risk-related, etc. assessments.
- May be unique to the collection/record/media type and institution-specific.

Essential Characteristics -

Approaches have been previously defined for photographic reformatting:

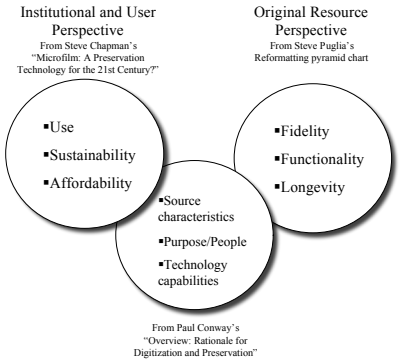
- Text on Microfilm - imaging approaches for maintaining text legibility, including legibility of smallest significant character
- Photographs - NARA Photographic Duplication Specifications for historic negatives

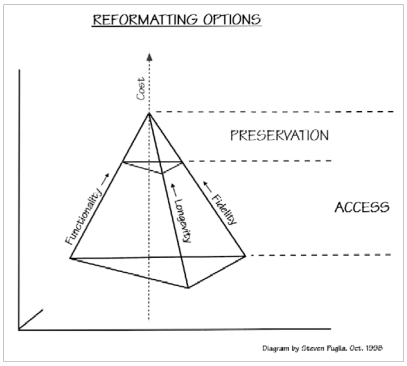
Technical Issues - Duplicating Still Photo Negatives:

- Method of producing duplicates.
- Tone reproduction approach.
- Objective means of evaluating tone reproduction.
- Variability.
- Choosing an appropriate film.

Considerations for Reformatting:

- Archival considerations
- Costs
- Reproduction quality
- Stability of imaging materials
- Ease of distribution





Goals:

- Produce consistent, high-quality digital objects and related metadata.
- Facilitate the long-term management and preservation of the digital resources.

Use appropriate standards whenever possible for:

- Metadata of all types
- File formats
- Approaches to digitizing physical collections
- Etc.

Collect and create as much metadata as possible:

- Descriptive / Discovery
- Administrative
- Technical
- Preservation
- Behavior / Structural

Our perspective has been-

- Standardize your digital objects, just like your metadata.
- Better to define consistent approaches.
- Treat large batches of images, or other digital objects, in the same way.

Standardization will promote ease of management and lower costs to maintain and preserve digital data/objects/records.

•Consciously capture as many characteristics and as much information as you think appropriate to define the original resources.

•Perform a cost-benefit analysis to determine the most cost-effective approach to reformatting.

Digitizing guidelines exist - still work to be done:

- •Textual records - original paper records and microfilm
- •Still photographs
- •Audio recordings
- •Working on video recordings and motion pictures

Digitizing Guidelines:

- •Benchmark for Faithful Digital Reproductions of Monographs and Serials
- •NARA's 2004 Technical Guidelines for Digitizing Archival Materials for Electronic Access
- •Federal Agencies Digitization Guidelines Initiative (FADGI)

NARA Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images, June 2004

- •Digital Image Capture
- •Minimum Metadata
- •File Formats, Naming, and Storage
- •Quality Control and Assurance

Not Addressed - in 2004 Tech Guide:

- •Project scope.
- •Selection
- •Preparation
- •Descriptive systems
- •Project management
- •Access to digital resources
- •Legal issues
- •IT infrastructure
- •Project assessment
- •Digital preservation

Metadata:

- Descriptive
- Administrative
- Rights
- Technical
- Structural
- Behavior
- Preservation
- Image Quality Assessment
- Records Management / Record Keeping
- Tracking
- Meta - Metadata

2004 Technical Guidelines - continued:

- Reference Targets –
  - Scale and dimensional
  - Tone and color reproduction
- Aimpoints – guidelines only, there are exceptions
- Image Processing – sample workflow
- Digitizing Specifications for Record Types
  - Requirements tables
- Quality Control

Metadata:

- NISO - Understanding Metadata

Cataloging, indexing, description, etc.

Many definitions

- "Data about data"
- In most cases, the data itself *is* as important as the digital resources
- Information necessary to identify, locate, manage, access, use, and preserve a digital resource of any media type or format
- Essential part of any digitization project

Metadata characteristics

- Content
- Consistency
- Coherence
- Context
- Conformance

Metadata characteristics

- Content
  - *Level of cataloging or description*
    Appropriate granularity for intended use or scope of digital project?
  - *Origin of data*
    Automatically or manually generated?
  - *Control*
    Appropriate standards followed and controlled vocabularies used?

Metadata characteristics

- Content
  - Data content standards
    - Structure: A set of rules for applying, using, and formatting values within fields
    - Examples: AACR2, DACS, CCO, …

Metadata characteristics

- Consistency
  - Follow the same practice for all related metadata records
  - Use consistent values for ease of search and discovery across metadata records
- Coherence
  - Use repeatable values when possible
  - Question intended use of field and relevance/appropriateness of value

Metadata characteristics

- Context
  - What information is locally important?
  - What information is shareable across different systems and audiences?
  - What is the intended use of the metadata?
  - Different types of metadata will be used in different contexts and systems for different purposes.

Metadata characteristics

- Conformance
  - Follow standards whenever possible to enhance interoperability and consistency in structure and content of metadata
  - Data structure standards (or schemas)
    - DC, MARC, MODS, VRA, …
  - Controlled vocabularies
    - AAT, LCSH, TGN, Mimetypes, …
  - Encoding standards
    - XML, HTML, Relational database, …

"Introduction to Metadata"

Edited by Murtha Baca
The Getty Research Institute

- Setting the Stage
- Practical Principles for Metadata Creation and Maintenance

Some uses of metadata

- Facilitate search and browsing
- Online display
- Navigation of multi-page or multi-part documents
- Identification and location of resources
- Long-term management of digital objects
- Format description and profiles
- Production, workflow, tracking
- Preservation
- Use and reproduction information
- Promote interoperability

Metadata types

- Descriptive
  - Indexing
- Technical
- Structural
- Administrative
  - Rights
  - Preservation
  - Workflow and Job Tracking

What are we documenting?

- The original resource
- The digital resource
- The process (as part of the resource's provenance to verify that the digital version is an accurate and authentic representation of the original)
- Changes to the digital resource over time
- The content regardless of format

When do you create metadata?

- Prior to digitizing - most often, descriptive metadata only
- During digitizing process - most of this metadata may only be of limited use
- During workflow - documents process, but how accurate is it?
- Post Capture/Digitization - sometime it may be easier to create metadata after conversion while looking at the digital resources
- Upon ingest into systems - extraction from digital file; identification, validation, and characterization; creation of checksums

Assessment for metadata creation

- Is there any existing metadata?
- Is it accurate and complete? In what format? Does it follow any standards?
- How much information about the resource is needed?
- At what level will it be realistic to record metadata (i.e., series, item, etc.)?
- What are the number and volume of items?
- Are they homogeneous?
- How are they organized?

Assessment for metadata creation

- Who will create and use the metadata (systems, humans)?
- How will metadata be captured - manually, automatically?
- At what points in the workflow will metadata be captured?
- Where will the metadata be stored? In what format?
- Does the metadata need to be carried forward? What is the intended use?

Levels of Metadata

- Metadata may exist at any level of aggregation
- May exist for a batch of digital files or for each individual file
- Levels appropriate to the specific project need to be defined upon project start-up

Levels of Metadata

- Need to ask: what is the minimum complement of metadata we can get by with?
- Will likely vary by type of project, i.e., large project vs. exhibit

Where do you create and store metadata?

- In digital repositories, digital asset management systems, or database-driven systems - fielded character data or XML
- In spreadsheets, XML documents (METS)
- Mapped from fields in other systems, databases, files
- Directly in the digital file itself (i.e., header tags and/or XML)

Digital Preservation

Preservation (IFLA):

- Preservation includes all the managerial and financial considerations including storage and accommodation provisions, staffing levels, policies, techniques, and methods involved in preserving library and archive materials and the information contained in them.

Preservation is a long-term management process addressing the identification and mitigation of the risks for loss of information, and for appropriate collections and records includes protection of the original physical form.

Digital preservation refers to a series of managed activities designed to prevent obsolescence and to maintain data integrity.

Digital objects are not preserved unless they are stored in a digital repository.

OCLC

Association for Library Collections and Technical Services definition-

- •Digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change.
- •Digital preservation applies to both born digital and reformatted content.

•Digital preservation policies document an organization's commitment to preserve digital content for future use; specify file formats to be preserved and the level of preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information.

Systems Perspective:

Managing and preserving digital data/objects/records is different than managing and preserving physical records.

- •It is about risk management.
- •It is about creating a managed environment – the need to be proactive, not just reactive.

Digital-

- •Everyone still trying to determine all that will be necessary to preserve digital data over the long term.
- •Will need to preserve digitally created materials using digital technology.
- •Media reversion is not an option for many types of records, will lose functionality and then it is no longer the record.

Stephen Chapman, "What is Digital Preservation"
http://www.oclc.org/news/events/presentations/2001/preservation/chapman.htm

Digital repository - a repository "…is understood to mean any organization or system charged with the task of preserving information over the long term and making it accessible to a specified class of users…"

Chapman Continued-

Preservation Obligation: Guard against obsolescence-

- •Images become incompatible with associated applications.
- •Images (and associated applications) become incompatible with current use requirements.
- •Preservation strategies strive to manage both technological compatibility and user expectations.
- •Fiscal obsolescence could prove to be one of the biggest problems.

Chapman Continued-

Maintain readability and integrity?:

- •What constitutes "integrity" should be made explicit in repository policies and procedures.
- •Determining what is meaningful is challenging, but essential.
- •Content, content and representation, provenance, all three?

Chapman Continued-

Repository is more than storage:

- •Submission agreements.
- •Storage systems to manage the digital objects (data).
- •Database to manage administrative information (metadata).
- •Data managers, Administrators.
- •Preservation policies/procedures.

Open Archival Information System (OAIS):

NASA has developed a reference model that defines terms for general functions of a digital repository used for the long-term preservation of digital data, but does not provide any implementation details.

http://nost.gsfc.nasa.gov/isoas/ref_model.html

Functional Overview - OAIS

- •Ingest
- •Archival Storage
- •Data Management
- •Administration
- •Preservation Planning
- •Access
- •Common Services

Potential actions and techniques for preserving digital data may include:

- •Maintaining obsolete systems
- •Data integrity / verification - use checksums
- •Refreshment – data and media
- •Migration – media and formats
- •Normalization
- •Transformation
- •Emulation
- •Digital archeology
- •New/future technical and conceptual approaches

Build appropriate information technology infrastructure, including well designed and robust IT systems.

For digital reformatting, need at least three tiers of information technology infrastructure-

- Digital Repository
- Shared Work Environment
- Digital Reformatting Lab

At all levels - need to do the routine IT things to mitigate risk of data loss.

Range of IT issues:

- Data validation and integrity
- Records integrity and provenance
- Data and system security
- Metadata (all types)
- Hardware and software
- Policies
- Staff

Effective IT procedures exist for the short-term management of electronic records and digital information-

- Not always followed.
- Not always as easy or as inexpensive as advertised.
- We have been sold on the promise of the technology, but rarely acknowledge the downsides.

Emphasize data security, including back-ups, distributed storage of multiple copies, etc., to prevent catastrophic loss of digital resources.

Digital Object Types Produced:

Dynamic -

- •Motion Pictures: time-based sequence of high resolution and high bit raster images synced to pulse code modulation sound recording
- •Video: time-based sequence of raster images synced to pulse code modulation sound recording
- •Audio: time-based pulse code modulation sound recording

Static -

- •Still Photographs: raster image
- •Textual Records: raster image

Increasing complexity for digital objects.

Types of Copies Produced:

- •Master Copies - warrant a level of effort to manage and maintain
  - ▪Preservation Master Files
  - ▪Production Master Files
  - ▪Reference Master Files
- •Distribution Copies - managed as part of delivery and access environment
  - ▪Derivative Files

Another perspective on digital preservation - NLNZ recommends creating an Institutional Technical Profile that defines specifically what your organization is prepared to preserve.

"The Costs of Digital Imaging Projects"

First presented (EMG Session of the AIC Annual Meeting) and published (DigiNews) in 1999.

Costs:

Selection
Preparation
Cataloging / Description / Indexing
Preservation / Conservation
Production of Intermediates
Digitization
Quality Control- Images and data
Network Infrastructure
On-Going Maintenance of Images
On-Going Maintenance of Data

Overall Average Costs:

- •Selected projects
- •NARA's Electronic Access Project
- •Published costs
- •Projected costs (LC/Ameritech rounds one and three)

On average:

- 1/3 the cost is digital conversion
- slightly less than 1/3 the cost is cataloging / description / indexing
- slightly more than 1/3 the cost is administrative / QC / etc.

Adjusted Projections (per image):

Total-	$17.65 [23.25]	range- $1.85 to $42.45
Digitizing-	$6.15	range- $0.25 to $16.65
	34% [32%]	5.5% to 80%
Cat./Desc./Ind.-	$7.00	range- $0.75 to $17.25
	31% [29%]	3.5% to 55%
Admin./QC-	$10.10	range- $0.45 to $28.15
	41% [39%]	19% to 78%

Single Items (per page):

Total-	$29.55 [32.90]	range- $23.10 to $35.80
Digitizing-	$5.30	range- $1.90 to $8.00
	15%	5.5% to 27%
Cat./Desc./Ind.-	$10.40	range- $5.75 to $12.85
	40%	16% to 56%
Admin./QC-	$17.20	range- $7.60 to $28.15
	45%	28% to 79%

Mixed Collections (per item):

Total-	$24.45 [31.35]	range- $3.25 to $40.50
Digitizing-	$9.35	range- $3.45 to $16.50
	37% [34%]	15% to 53%
Cat./Desc./Ind.-	$10.60	range- $2.85 to $17.25
	32% [30%]	14% to 46%
Admin./QC-	$11.40	range- $4.50 to $21.55
	39% [36%]	19% to 42%

Photo Collections (per photo):

Total-	$19.30 [26.90]	range- $5.20 to $42.45
Digitizing-	$7.60	range- $2.30 to $16.65
	28% [27%]	11% to 58%
Cat./Desc./Ind.-	$5.85	range- $4.85 to $6.45
	28% [27%]	15% to 34%
Admin./QC-	$13.45	range- $3.35 to $24.65
	48% [46%]	24% to 58%

Re-Keying Text (per page):

Total-	$8.80 [11.40]	range- limited data
Digitizing-	$3.50	range- $2.55 to $5.00
	31% [30%]	limited data
Cat./Desc./Ind.-	$4.00	range- $2.35 to $5.70
	18% [18%]	8.5% to 27%
Admin./QC-	$3.90	range- limited data
	53% [52%]	44% to 62%

Multi-Page (per page):

Total-	$8.35 [13.45]	range- $4.60 to $14.40
Digitizing-	$4.30	range- $2.10 to $6.10
	49% [47%]	14% to 80%
Cat./Desc./Ind.-	$5.60	range- $1.50 to $11.10
	26% [25%]	4% to 50%
Admin./QC-	$3.55	range- $1.35 to $6.90
	29% [28%]	16% to 48%

OCR (per page):

Total-	$4.40 [4.05]	range- $1.85 to $7.65
Digitizing-	$1.20	range- $0.25 to $3.60
	42% [38%]	36% to 47%
Cat./Desc./Ind.-	$1.45	range- $0.75 to $2.40
	35% [31%]	30% to 40%
Admin./QC-	$1.40	range- $0.40 to $2.10
	34% [31%]	21% to 59%

National Digital Library Program, Library of Congress:

- •Goal - produce 5 million digital images over 5 years for an approximate cost of $60 million
- •An estimated $12 per image (51% of median average overall cost)
- •At one point, NDL had a staff of approximately 85 people

NDL 2001 Annual Review: 7.5 million images

Assuming for online collections -

- •50% have 2 each images or versions
- •25% have 3 each images or versions
- •25% have 4 each images or versions

7.5 million images represents -

- •Approximately 3 million unique items and/or images

- •NDL website cites $60 million over 5 years, 1996 to 2000.
- •$60 million ÷ 3 million items and/or images = $20 per image (86% of median average overall cost)
- •Cost does not include contributions of LC/Ameritech awardees, 33 institutions. These projects represent 20% of the online collections on American Memory site. LC/Ameritech awarded only $1.75 million over three years of grant program.

NDL Expenses- approx. $43 million over 5 years:

Year	Personnel	Digitization and Services	Professional and Consulting Services
1997	43%	22%	12%
1998	43%	29%	9%
1999	43%	29%	10%
2000	53%	24%	15%
2001	12%	30%	47%
Average-	46%	27%	18%

NDL Annual Reviews- lcweb.loc.gov/fsd/fin/

National Yiddish Book Center:

- $3.5 million project digitized 12,000 books or $292 per book

Nina Thayer, "Books go online, Yiddish goes digital" in *The Gantseh Megillah*, Associated Press, 2002.
www.pass.to/newsletter/0602YiddishGoesDigital.htm

(12% of average overall cost for multi-page for average book)

Corbis - Bettman Archive

Began scanning in 1996 at a cost of $20 per photograph (104% of average overall cost for photos).

Stopped systematic digitization after scanning approximately 225,000 photographs, out of approximately 7.5 million original images (11 million total).

Corbis has approximately 2.1 million digital images online for licensing, out of 65 million total.

Denver Public Library: Western History / Genealogy Department

- Cost to digitize and catalog a photograph-

  $25 to $35 (130% to 181% of average overall cost for photos).

- Includes preparation, research, cataloging, and scanning.
- Does not include selection, curatorial decisions, equipment purchases and upgrades, or administrative and supervision costs.

http://photoswest.org/faqs3.htm

Denver Public Library: Western History / Genealogy Department

- Pictures processed per day:
  - Preparation of photographs- 100 per day
  - Research and creation of MARC record- 20 per day
  - Cataloging- 40 per day
  - Scanning- 55 grayscale images or 40 color images per day

http://photoswest.org/faqs3.htm

Average of 52 photos per day (116% of overall average).

Boulder Public Library: Carnegie Historical Images:

- About $15 per image (78% of average overall cost for photos)

"Appendices for Austin History Center Business Case for Digitizing Photographs, Best Practices Questions Used in Gathering Data"
www.gslis.utexas.edu/~ssoy/pubs/ahcdigital/appendices.htm

U.S. Steel Photograph Project

- 2200 images from the U.S. Steel Gary Works Photograph Collection held by the Calumet Regional Archive at IU Northwest.
- TOTAL PROJECT COSTS: $40,730
- $18.51 per image

  (96% of overall average for photos)
- Completed in 18 months

Courtesy K. Brancolini

"Costs and Funding" by Robin Crumrin
ALI Digital Library Workshop: October 2, 2003

Charles Cushman Slide Collection

- 15,000 archival color slides, the work of amateur photographer Charles Cushman, held by the University Archives at IU Bloomington.
- TOTAL PROJECT COSTS: $301,937
- $20.13 per image

(104% of overall average for photos)

- Scheduled for completion in 30 months

"Costs and Funding" by Robin Crumrin

ALI Digital Library Workshop: October 2, 2003

Russian Periodical Index Project

- Converting a 20-year run of a Russian periodical index to an online version
- 234,000 bitonal page images (text only)
- Outsourced to (Northern Micrographics):
- CREATION OF PAGE IMAGES (not the entire project): $32,760
- $ 0.14 per page image
- Completed in 8 months

"Costs and Funding" by Robin Crumrin

ALI Digital Library Workshop: October 2, 2003

On-Going Costs

Plan for the on-going costs from the beginning of the project.

Cost for minimal maintenance of one set of the master image files (off-line) and access files (on-line) during the first 10 years likely to be 50% to 100% of initial investment.

Cost for maintaining image files in large-scale automated digital repository during the first 10 years likely to be 10% to 25% of initial investment.

Cost to install, staff and maintain infrastructure and the digital data for 1st 10 years is up to 5 times the initial investment.

In the IT world, typically the full lifecycle cost (for a 7 to 10 year lifecycle) for a system is up to 10 times the cost of development.

Cost Reduction

Sustainability

Blue Ribbon Task Force on Sustainable Digital Preservation and Access

Conclusions:

- In today's world - reformatting digitization plays an important role in libraries, archives and museums.
- Digitization can be complex, but can be done successfully.
- Everyone is making progress, but there is still a lot of work to do in regards to digital preservation and long-term sustainability.