

KU Digital Initiatives – Recommended Standards and Best Practices for Digital Projects

January 2003

This document outlines a set of "best practices" for creating digital representations of scholarly materials at the University of Kansas. The recommendations here focus on the initial stages of capturing digital images and metadata, and do not cover other important issues such as systems architecture, search and retrieval, interoperability, and longevity (*for information on these issues, please contact the KU Digital Library Initiatives office*). The information covered focuses primarily on reformatting existing works (such as handwritten manuscripts, typescript works on paper, bound volumes, slides, or photographs) into digital formats, but can readily be applied to new, "born digital" materials as well.

These recommendations are directed towards projects creating digital objects which are expected to persist for a long period of time and easily migrate to new systems and technologies. Projects creating digital objects with anticipated relatively short life-spans may choose to follow a less rigorous set of guidelines. However, this should be done with the understanding that the resulting collection, which may become a valuable resource, may not be acceptable for long term archival storage and the digitization process may need to be redone at a later date.

Because image capture capabilities are still changing, this document covers practices felt to be fairly universal, and usable for many years to come. This includes the notion of masters and derivatives, information about image quality and file formats, and metadata capture / creation.

Digital Objects

Principles of "Good" Digital Objects

- **A good digital object will be produced in a way that ensures it supports collection priorities.**
- **A good object is persistent.** That is, it will be the intention of some known individual or institution that the good object remain accessible over time despite changing technologies.
- **A good object is digitized in a format that supports intended current and likely future use or that support the development of access copies that support those uses.** Consequently, a good object is exchangeable across platforms, broadly accessible, and will either be digitized according to a recognized standard or best practice or deviate from standards and practices only for well documented reasons.
- **A good object will be named with a persistent, unique identifier** that conforms to a well-documented scheme. It will not be named with reference to its absolute filename or address (e.g. as with URLs and other Internet addresses) as filenames and addresses have a tendency to change. Rather, the filename's location will be resolvable with reference to its identifier.
- **A good object can be authenticated in at least two senses.** First, a user should be able to determine the object's origins, structure, and developmental history (version, etc.). Second, a user should be able to determine that the object is what it purports to be.
- **A good object will have, and be associated with, metadata.** All good objects will have descriptive and administrative metadata. Some will have metadata that supplies information

about their external relationships to other objects (e.g. the structural metadata that determines how page images from a digitally reformatted book relate to one another in some sequence).

Types of Digital Objects

In general, there are two kinds of digital objects:

- those produced as **surrogates** for information objects that exist in some analog format (e.g. as books, manuscripts, museum artifacts, audio or video tapes, etc.), and
- those that are **born digital**, i.e., that are produced originally in machine-readable form (scientific databases, sensory data, digital photographs, etc.).

A good object created as a surrogate will be considered a faithful facsimile of the artifact.

When discussing digital objects, it is often useful to distinguish between **master or preservation copies** and **access or use copies**:

- Masters are typically the highest quality versions that the production technique allows
- Use or access copies are derivatives created for specific uses, distribution scenarios, or users.

For example, a master copy of a digitally reformatted 35mm slide might be an uncompressed, 18 megabyte, TIFF file, captured in 24-bit color, at a resolution of 600 dpi. The access or derivative copy of this might be a 150 KB, JPEG image derived from the TIFF file, which will allow a reasonable download time for the average Web-based user.

Why worry about standards for “good” digital objects?

- **Reduce Risk:** Setting minimum level benchmarks can reduce the risk involved in producing and maintaining digital objects while inspiring confidence in and encouraging their use.
- **Future Investment:** Because good objects will be considered capable of meeting known current and likely future needs, organizations can invest in their creation secure in the knowledge that they will not be forced to re-create the objects at some future date even as production techniques improve.
- **User Confidence:** Good objects inspire confidence because they have a minimum level of well-known and consistent properties, and will support a variety of known uses.
- **Effective Practices:** By building consensus around the characteristics of good objects, organizations can more effectively:
 - write contracts with vendors and compare vendors' prices for object conversion
 - commit to making good objects accessible over the longer term
 - implement efficient data creation processes and access systems
 - instill confidence in users regarding object quality and support
 - define and narrow preservation options for migration or emulation of good objects

Digital Object Format Standards / Best Practices

In almost every case, there is a direct correlation between the production quality of a digitized object and the readiness and flexibility with which that object may be migrated across platforms. As a result, the digitization of objects at the highest affordable quality can pay off in the long run as the objects are more useful and more accessible over the longer term. However, not all objects require such investment. Every digitization project needs to determine the value of the digitized objects themselves and make appropriate decisions about persistence and interoperability.

Common formats for original materials and recommended formats for digital conversion are presented *Tables 1 and 2*.

Metadata

Metadata is an important component of digital objects, as it supports the discovery, use, storage and migration of these objects over time. Metadata must be collected and associated with each digital object as part of the object creation process. Three types of metadata are associated with digital objects:

- **Descriptive Metadata** is used in the discovery and identification of an object. Examples include Dublin Core, VRA, and MARC records. Additionally, descriptive metadata for digital objects applies to information on the full collection of files associated with the digital object and their relationships to one another.
- **Structural Metadata** is used to display and navigate a particular object for a user and includes the information on the internal organization of that object (e.g., a book may have an introduction, chapters, pages and an index).
- **Administrative Metadata** represents the management information for this object, including the date it was created, its content file format (TIFF, JPEG, GIF, etc.), scanning resolutions used, rights information, etc.

In addition to the data structures used to convey metadata, there are a number of standard resources available for determining the **semantics and syntax** of the metadata content. These are usually application- or data format- community specific and include resources such as:

- AACR2 (Anglo-American Cataloging Rules)
- LCSH, (Library of Congress Subject Headings)
- AAT (Art and Architecture Thesaurus) http://shiva.pub.getty.edu/aat_browser/
- CDWA (Categories for the Description of Works of Art) <http://www.getty.edu/gri/standard/cdwa/>
- TGN (Getty Thesaurus of Geographic Names) http://shiva.pub.getty.edu/tgn_browser/
- ULAN (Union List of Artist Names) http://shiva.pub.getty.edu/ulan_browser/.

For more detailed information on recommended metadata practices at KU, see *Digital Library Metadata: A report from the Digital Library Metadata Working Group To the Digital Library Executive Group*, August 31, 2001, http://kudiglib.ku.edu/projects/wgs/metadata_wg/MetadataReportFinal.doc, or contact the KU Digital Library Initiatives office.

Legal Considerations - Copyright

The University supports the production of intellectual property by faculty and students for the benefit of the institution and society. All users and creators of digital information have a personal responsibility to recognize and honor the intellectual property of others.

Some basic background information and guidelines for the use of copyrighted material in the educational environment can be found at <http://www.ku.edu/~vcinfo/Copyright/copyright.htm>. For more information or assistance in answering copyright questions, contact the Vice Provost for Information Services Office at 864-4999 or copyright@ku.edu.

Summary of General Recommendations

- Think about users (and potential users), uses, and type of material/collection
- Scan at the highest quality you can possibly justify based on potential users/uses/material. Err on the side of quality.

- Do not let today's delivery limitations influence your scanning file sizes; understand the difference between digital masters and derivative files used for delivery.
- Many documents which appear to be bitonal actually are better represented with grayscale scans; some documents that appear to be grayscale are better represented in color
- Include grayscale target, standard color patches, and ruler in the scan.
- Use objective measurements to determine scanner settings (do NOT attempt to make the image good on your particular monitor or use image processing to color correct).
- Don't use compression for digital masters.
- Store in a common (standardized) file format.
- Capture as much metadata as is reasonably possible (including metadata about the scanning / creation process itself). Metadata should follow accepted national and local standards whenever possible. Non-conforming metadata formats should be well justified and documented.

Selected Resources

California Digital Library, various publications found at <http://www.cdlib.org/about/publications/>

Gilheany, Steve, *Course in Document Imaging – Document Management*, <http://www.archivebuilders.com/whitepapers/index.html>

History Data Service UK Data Archive, University of Essex, information on creating digital resources from historical sources, <http://hds.essex.ac.uk/create.asp>

IMLS, *A Framework of Guidelines for Building Good Digital Collections*, 2001, <http://www.ims.gov/pubs/forumframework.htm>

Kenney, Anne R. and Oya Y. Rieger, *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, Mountain View, CA:Research Libraries Group, 2000, 189 pgs. Online tutorial at <http://www.library.cornell.edu/preservation/tutorial/contents.html>

Kenney, Anne R. and Oya Y. Rieger, Co-Chairs, *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, Cornell University Library

Lee, Stuart D., *Digital Imaging: A Practical Handbook*, NY:Neal-Schuman Publishers, Inc., 2001, 194 pgs.

Library Preservation Office, Harvard University, *Digitization Resources*, <http://preserve.harvard.edu/resources/digital.html>, November, 2002

Sitts, Maxine K., ed., *Handbook for Digital Projects: A Management Tool for Preservation and Access*, Northeast Document Conservation Center, 2000, <http://www.nedcc.org/digital/dighome.htm>

Contact Information for Additional Information

Beth Forrest Warner, Director, Digital Library Initiatives, University of Kansas, (785) 864-4999, bwarner@ku.edu, <http://kudiglib.ku.edu>

TABLE 1. FORMAT STANDARDS

DATA TYPE	APPLICATIONS	FORMATS: MASTERS	FORMATS: ACCESS COPIES	GUIDELINES AND REFERENCES
Alphanumeric data	Flat files; hierarchical or relational datasets.	Comma-delimited ASCII, or portable format files recognized as de facto standards (e.g. SAS and SPSS) with enough metadata to distinguish tables, rows, columns, etc.	Same	For social science and historical datasets, see <i>Guide to Social Science Data Preparation and Archiving</i> (ICPSR 2000) http://www.icpsr.umich.edu/ACCESS/dpm.html , and <i>Digitizing history, a guide to creating electronic resources from historic documents</i> (HDS, 1999) http://hds.essex.ac.uk/create.asp
	Encoded texts for networked presentation and exchange of text-based information	SGML XML	Same	Use documented DTD's or schema
	Encoded texts for literary and linguistic content analysis	SGML XML	Same	Text Encoding Initiative (TEI) http://www.tei.org . <i>Creating and documenting electronic texts</i> (OTA, 1999) and http://hds.essex.ac.uk/create.asp <i>TEI text encoding in Libraries: Guidelines for Best Practice</i> (DLF, 1999) http://www.diglib.org/standards/tei.htm
Image data (raster graphics): bitonal, grayscale and color images of pictures, documents, maps, photographs, slides, etc.	Book or serial publication prepared as preservation digital master or access surrogate for source	See Table 2	Will vary depending on use. Preferred formats include: JPEG GIF PDF SID	Anne R. Kenney, Oya Y. Rieger, et al, <i>Report of the Digital Preservation Policy Working Group on Establishing a Central Depository for Preserving Digital Image Collections</i> (March 2001) at http://www.library.cornell.edu/preservation/IMLS/image_deposit_guidelines.pdf Library of Congress, <i>The Preservation Digital Reformatting Program: Image Specifications</i> (September 2001). The most recent consensus is available in Draft benchmark for digital reproductions of book and serial publications (DLF).

DATA TYPE	APPLICATIONS	FORMATS: MASTERS	FORMATS: ACCESS COPIES	GUIDELINES AND REFERENCES
				2001), at http://www.diglib.org/standards/draftbmark.htm An example of one institution's local benchmarks: California Digital Library. <i>Digital Object Format Standards</i> . http://www.cdlib.org/about/publications/
Scalable image data (vector graphics): presentations, creative graphics, computer-aided designs, clip art, line drawings, 3-D models, maps	maps, herbarium specimens	MrSid from LizardTech becoming de facto standard although proprietary		
Multimedia	GIS	GIS often combines data in multiple formats: GPS, alphanumeric data (e.g. as required to record co-ordinate data), vector and raster graphics (e.g. to represent maps)		<i>GIS. A guide to good practice</i> (ADS, 1998) http://ads.ahds.ac.uk/project/goodguides/gis/index.html
Audio	music audio	IFF or AIFF wav 44.1 KHz sample rate 16-bit	RealAudio wav mp3	A brief technical introduction to <i>Digital Audio</i> by the National Library of Canada http://www.nlc-bnc.ca/9/1/p1-248-e.html Harvard University Library Digital Initiative <i>Audio Reformating</i> http://hul.harvard.edu/ldi/html/reformatting_audio.html . Currently the site has links to industry standards and will include project guidelines in the future. <i>Sound Practice: A Report on the Best Practices for Digital Sound Meeting</i> , 16 January 2001 at the Library of Congress http://www.rlg.org/preserv/diginews/diginews5-2.html#feature3

DATA TYPE	APPLICATIONS	FORMATS: MASTERS	FORMATS: ACCESS COPIES	GUIDELINES AND REFERENCES
Audio	spoken word (e.g. oral histories)	No universally accepted standards yet. IFF or AIFF wav 44.1KHz sample rate 16-bit	RealAudio wav mp3	See music audio above. National Gallery of the Spoken Word http://www.historicalvoices.org/papers/sounds.rtf Old Dominion University on-line Digital Training Workshop, 2000 http://www.lib.odu.edu/services/dcenter/dtw2000/index.html
Video			<i>Moderate-resolution downloadable files:</i> Image size: 320x240 pixels Frame rate: 30 fps Data rate: ca. 1.2 MB/S(ca. 150KB/S) Compression: MPEG-1 Format: mpg <i>Low-resolution downloadable files:</i> Image size: 160x120 pixels Color depth: 24 bits/pixel Data rate: ca. 100 KB/S Format: QuickTime (Apple Computer format) File ext: mov <i>Streaming video:</i> RealVideo	Fleischhauer, Carl, <i>Digital Formats for Content Reproductions</i> , National Digital Library Program, Library of Congress, July 13, 1998 http://memory.loc.gov/ammem/formats.html

Table 2: DIGITAL MASTER IMAGE FILES – RECOMMENDED IMAGING REQUIREMENTS

DOCUMENT TYPE	RESOLUTION	BIT DEPTH	ENHANCEMENTS ALLOWED	FILE FORMAT	COMPRESSION
Printed Text	600 dpi +	Bitonal	Sharpening, descreening, cropping, de-skewing, and de-specking	TIFF 5 TIFF 6	Lossless compression (e.g. ITU-G4)
Rare / damaged printed text	400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Book Illustrations	400 dpi 600 dpi with enhancement	8-gray or 24-color bitonal	Contrast stretching Minimal adjustment for tone and color Descreen / rescreen, sharpen	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Manuscripts	300-500 dpi	8-gray or 24-color, if color present in original	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Maps & other oversized items	300-400 dpi	8-gray or 24-color	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Graphic art	400-600 dpi	8-bit / channel internal reduction	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Photographic prints	400 dpi	8-bit / channel internal reduction	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Works of art on paper	400 dpi	8-bit / channel internal reduction	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Transparencies	4000-5000 on long end or 400 dpi on output > 8"x10"	8-bit / channel internal reduction	Contrast stretching Minimal adjustment for tone and color	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. LZW)
Microfilm	600 dpi blown back to original size 300-400 dpi blown back to original size	Bitonal 8-gray	Sharpening, descreening, cropping, de-skewing, and de-specking	TIFF 5 TIFF 6	Uncompressed or lossless compression (e.g. ITU-G4, LZW)

+ Although 600 dpi 1-bit is a defacto standard for printed text, a comparable or richer text file may be produced in grayscale at 400 dpi