5-1-2007

Digital Collections User Group Update and Responses to the Digital Vision Task Force Report

Consortium of Academic and Research Libraries in Illinois

Digital Collections User Group Update and Responses to Digital Vision Task Force Report

Submitted to the Board for Review
May 1, 2007

Members
Howard Carter, Southern Illinois University Carbondale
Jessica Gibson, CARLI
Denise Green, University of Illinois at Springfield
Katharine James, Illinois State University
Amy Maroso, CARLI
Michel Nguessan, Governors State University
Ellen Starkman, University of Illinois at Chicago
Claire Stewart, Northwestern University
Kerri Willette, School of the Art Institute
Suzanne Wilson; chair, Illinois Wesleyan University

Digital Collections User Group Update

The Digital Collections User Group (DCUG) was formed by the CARLI Board of Directors in the summer of 2006. Since this is a newly formed group and digital collections are a new initiative for CARLI, the group has spent a significant amount of time discussing a broad range of issues.

During these discussions, the need for immediate development in three areas was apparent. Presently, there are three subcommittees that are working in the following areas.

> Interface Subcommittee – This committee is looking at the public interface for the CARLI Digital Collections Website and identifying areas that need further development. This committee will be working on a redesign of the site and doing usability testing.
>
> Standards Subcommittee – This committee is developing standards and best practice recommendations for digital collections in the CARLI consortium.
>
> SASKIA Webinar/Tutorials Subcommittee – This committee is looking at ways to promote the use of SASKIA in the CARLI community. They are also developing a number of online tutorials for using CONTENTdm.

In addition to forming these working committees, the DCUG is working on a survey to send out to the CARLI community to help us determine what types of collections institutions are working on and what additional needs they may have (see Appendix A). We have tried to keep the survey general to include both CONTENTdm users and institutions that may be using other systems. Our goal is to send the survey out at the beginning of May and to start evaluating the results in June.

The general consensus among the DCUG is that we should be fostering as open an environment as possible to allow libraries to host a wide range of collections and to explore new partnerships, both within the organization and beyond. The first step to moving in this direction will be a revision of the existing Collection Development Policy. We could host a one day seminar, perhaps in the fall, to provide examples of what libraries are doing with collections and to help inspire ideas.

The DCUG is also interested in looking into different ways that we can allow others to access and search the metadata in our collections, such as; OAI, federated searching, etc. We would like to start thinking beyond just providing a local CARLI search and to explore ways to also push our information out into other systems to increase access to the CARLI digital collections.

Responses to the Digital Vision Taskforce Report

In March of 2007, the Digital Collections User Group (DCUG) received a list of issues from the Digital Vision Task Force (DVTF). We were asked to respond back to the Board of Directors by the end of April 2007.

The DCUG has discussed the issues that were sent to us from the DVTF. Two of the questions, number 1 and 6, feel very strategic. We did our best to respond with some options to both of these questions, but feel strongly that more discussion is necessary by the Board and possibly the CARLI community.

1.) Should CARLI become a steward of cultural materials and, if so, in what ways? What are the central issues regarding long-term storage and preservation of digital objects?

Response

The DCUG group feels that CARLI should not act as a steward of cultural materials. We defined "steward" as playing an active role in the collection and preservation of digital objects**.** The responsibility for collections and preservation of cultural materials should reside with the member libraries.

There are some precedents for contributed content, like Open Content Alliance, where the individual libraries are taking on the responsibility for archiving contributed files. Also, some libraries may already be involved in other collaboration for archiving, such as the Committee on Institutional Cooperation (CIC) efforts.

However, the group would like to see CARLI explore if there is a need in the consortium for long-term storage solutions. This could go beyond access level versions of digital collection objects. Could CARLI provide such a solution through the leasing of hard drive space? CARLI's stewardship role would be limited to maintaining the bit-level integrity of data stored on central CARLI servers. Overall responsibility for collections and preservation of cultural materials would still reside with the member libraries.

Regarding potential storage problems at CARLI institutions, we are doing a survey (see Appendix A) in early May that includes some questions to collect data on this topic.

If our survey shows that there is a need for a long-term storage solution, the DCUG recommends that the CARLI Board form a committee to further investigate this idea. This committee should also explore if and how any other consortia are addressing this issue.

Issues to address with long term storage:

- Hard drive/server space requirements
- Personnel resources
- Hardware/Software maintenance
- Accessibility
- Data backup
- Pricing / Fee Structure
- Procedures for submitting and accessing files

2.) Should CARLI set minimal standards for metadata and if so, what are they?

Response

Presently, there are only two required metadata elements for CARLI Collections, the title element is a CONTENTdm requirement and the collection name element is a CARLI requirement**.** The DCUG is making a recommendation to require a metadata element for a rights statement. This recommendation is being sent on to the Board for approval. As the number of CARLI collections grow, the DCUG could identify additional elements that should be required and not just recommended.

At this time, we are also working on a Metadata Best Practices document that will recommend metadata guidelines based on existing standards (see Appendix B and C).

3.) Should CARLI set other consortial standards and is so, what are they?

Response

The DCUG is still in the process of creating documentation that will define minimal scanning best practices for the CARLI community**.** These documents include guidelines for scanning images and text (see Appendix D and E). We hope to publish these documents soon and then begin working on creating minimal guidelines for audio and video.

In addition, we have created a document (http://www.carli.illinois.edu/mem-prod/contentdm/digresrcs.pdf) which identifies existing online resources that have best practices and recommendations in the following areas:

- General Guidelines and Standards
- Metadata Standards
- Imaging Guidelines and Standards
- Metadata Crosswalks

4.) Scribe scanning station at UIUC – pros/cons and recommendation?

Response

Our understanding is that the Scribe Scanning Station was recently purchased by UIUC and that they are just now in the process of working with it. So, UIUC needs some time to develop procedures and workflow for using this equipment. This is something CARLI should follow up with UIUC on for an update in six months.

We have also included a question in our survey to try to determine what the need is for scanning bound materials. This may help CARLI decide if this is something worth pursuing.

Listed below are some of the pros and cons we have identified for implementing a Scribe Scanning Station:

Pros	**Cons**
Cost savings	Increased storage demands
Time savings	Logistics
Allows libraries to digitize collections they might not otherwise digitize	Quality control and delivery
	Copyright issues
	Handling of rare and fragile materials
	Personnel issues

5.) Linking of servers is happening at the statewide level – what is our role?

Response

It is our understanding that presently the Illinois State Library is taking the lead with this project and CARLI is playing a supporting role. We encourage CARLI to continue being a part of these discussions and to participate in this endeavor.

In addition to that, we would also like to explore other avenues for us to make our collections as accessible as possible. Until now, we have been thinking in terms of users coming to our site and finding our collections, we also need to find other ways to make our collections more accessible.

Keeping in mind CARLI's limited staff resources, steps should be taken to ensure that metadata for centrally hosted collections can be accessed and analyzed using external tools. This may include installing and supporting a Z39.50 service, an XML gateway, and/or becoming an OAI data provider.

6.) Is there a role for CARLI in developing new models of scholarly communication, OR…
In helping educate faculty and students on CARLI campuses about scholarly communication?

Response

The DCUG does not think that CARLI should be taking a leading role in the discussions surrounding new models of scholarly communication. Other organizations like ARL and ACRL are providing leadership in these areas.

However, we do feel that because of CARLI's involvement in CONTENTdm and digital collections that we have a responsibility to foster awareness and educate CARLI campuses about scholarly communication.

There may be member libraries who would like to use CONTENTdm or other CARLI systems to offer campus services such as institutional repositories or online journals. CARLI can be a clearinghouse for best practices in this area, and house shared documentation addressing specific issues likely to arise when accepting submissions from community authors. These may include copyright education, rights and archiving agreements, strategies for supporting an editorial process, and recommendations for publishing tools that could be deployed instead of or in tandem with CONTENTdm, such as bepress, DSpace or the Open Journal System (OJS).

The above suggestions are worth more conversations by the Board and the CARLI community.

Summary

Since digital object management, presentation and preservation are such new initiatives and there are so many opportunities to consider, the DCUG appreciates the communication with the Board and being a part of these discussions. The DCUG is available for any clarification or feedback. Please let us know if we can be of any further assistance.

Appendix A

SURVEY QUESTIONS

1. Content Areas
    a. What are the major subject areas of the materials you plan to digitize in the next year or two?
    b. What types of content would you like to see included in a consortial digital library?
    c. What are the major subject areas of your current digital collections?
    d. Are you using a Federated Search system to search digital collections?
        i. If so, what system are you using?
        ii. What collections are you searching?

2. Collection Storage
    a. Are you currently using a Digital Collection Management System?
    b. If so, what system are you using?
    c. Do you forsee using CONTENTdm in your library?
    d. How are you storing your images for preservation?
    e. Is this system adequate to meet your long-term needs?

3. Collection Creation
    a. Equipment
        i. Does your library have any equipment for digitization beyond the standard flatbed scanner?
        ii. If so, what special purpose does this equipment fulfill?
        iii. Would you be willing to allow other CARLI libraries to send projects that need this equipment to your library for a fee?
    b. File Standards
        i. What types of digital objects does your library create?
            1. Still images
            2. 3D images
            3. Text files
            4. Video files
            5. Sound files
            6. Other
    c. Outsourcing
        i. Have your used a vendor to outsource any part of a digitization project?
        ii. Do you have a list of vendor's that you could recommend?
        iii. Do you have the need to outsource the scanning of materials?
            1. Bound
            2. Images
            3. Newspapers
            4. 3D
            5. Audio
            6. Video

d. Data Standards
    i. What metadata scheme is associated with your existing files? (Mark all that apply)
        1. Dublin Core
        2. VRA Core 3.0 or 4.0
        3. MARC
        4. MODS
        5. TEI
        6. EAD
        7. CDWA Lite
        8. None
        9. Other (please specify/describe)
    ii. What is the controlled vocabulary used with your existing files? (Mark all that apply)
        1. Library of Congress Subject Headings
        2. Thesauri for Graphic Materials
        3. Art & Architecture Thesaurus (Getty)
        4. Union List of Artist Names (Getty)
        5. Getty Thesaurus of Geographic Names (TGN)
        6. None
        7. Other, Specialized, Local or unique vocabulary (please specify)

4. Funding
    a. How are you currently funding your digital collection projects?
    b. Do you have local expertise for writing grants?
    c. Do you know where to look for grants to help fund digital projects?

5. Training
    a. What digital project training sessions have you or your staff attended?
    b. Have you attended any training that you would like to recommend to others?
    c. What additional training needs do you have?
        i. Project Planning
        ii. Grant Writing
        iii. Metadata Creation
        iv. Creating Collections
        v. Digitization
        vi. Other

Appendix B

Descriptive metadata: Recommended practice for CARLI CONTENTdm collections

v.1 April 24, 2007

Libraries managing digital collections in the CARLI CONTENTdm system may use any descriptive metadata scheme to describe objects, but Dublin Core is recommended as a minimum standard. Descriptive or bibliographic metadata, such as Dublin Core, VRA Core, and MARC, describe the content of an object. Other types of metadata provide information about the structure and arrangement of the object (structural metadata) or its lifecycle as a digital object (administrative or technical metadata). Structural and administrative metadata are not the subject of this document.

Dublin Core is recommended both because it is relatively simple to implement and because it is often used as an exchange and interoperability format. For example, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) support unqualified[1] Dublin Core for metadata exchange between organizations.

The chart on the following pages provides a basic overview of version 1.1 of the Dublin Core element set, and indicates which elements are recommended or required for digital collections in CARLI's CONTENTdm system. Whether the library uses Dublin Core or another descriptive standard, some consideration should also be given to controlled vocabulary for personal names, subjects, and other key access points. Vocabulary is discussed briefly in the attached element chart.

For additional information about metadata and recommended best practices:

Dublin Core Metadata Element Set, Version 1.1

<http://dublincore.org/documents/dces/>

Collaborative Digitization Program Dublin Core best practices:

<http://www.cdpheritage.org/cdp/documents/CDPDCMBP.pdf>

Introduction to Metadata, Getty Institute

<http://www.getty.edu/research/conducting_research/standards/intrometadata/>

CONTENTdm system documentation

[1] The Dublin Core standard includes optional element refinements, or qualifiers, that provide additional specificity. For example, the Dublin Core element "date" has the qualifiers "created" and "dateAccepted." Since protocols like OAI-PMH only support unqualified Dublin Core, the record must still be useful if the refinements are removed. Dublin Core refinements are discussed at length on the DCMI site and in the Collaborative Digitization Program Dublin Core Best Practices document.

Appendix C

	CARLI Dublin Core implementation recommendations v.1			
Term or element	**Notes/description/details**	**CARLI status (required, recommended, optional)**	**Common vocabularies or encoding schemes**	**Example (s)**
Title	A name given to the resource. (Mandatory for Collaborative Digitization Program)	Required		
Creator	A person or entity primarily responsible for making the content of the resource. (Mandatory for Collaborative Digitization Program if available)	Recommended	Library of Congress Name Authority File	
Subject	The topic of the content of the resource. (Mandatory for Collaborative Digitization Program)	Recommended	Library of Congress Subject Headings, Art and Architecture Thesaurus, Thesaurus for Graphic Materials	
Description	An account of the content of the resource. (Mandatory for Collaborative Digitization Program) May include table of contents or abstract.	Recommended		
Publisher	An entity responsible for making the resource available.	Optional		
Contributor	An entity responsible for making contributions to the content of the resource.	Optional		
Date	A date associated with an event in the life cycle of the resource. Best practice is to input the date in the format YYYY-MM-DD according to the W3C-DTF scheme. (Date Digital and Date Original Mandatory for Collaborative Digitization Program)	Recommended	W3C-DTF	
Type	The nature or genre of the content of the resource.	Recommended	DCMI Type vocabulary	Image, MovingImage, Text
Format	The physical or digital manifestation of the resource. (Mandatory for Collaborative Digitization Program)	Recommended	Internet Media Type (IMT)	jpeg, jp2, tiff
Identifier	An unambiguous reference to the resource within a given context. (Mandatory for Collaborative Digitization Program)	Required (do we let this repeat?)	URI	

Source	A reference to a resource from which the present resource is derived. Source is not used to describe the nature of the relationship (see the Relation element), but to provide a pointer to the resource itself.	Optional	URI	
Language	A language of the intellectual content of the resource.	Optional	ISO639-2, RFC1766, RFC3066.	
Relation	A reference to a related resource. Extensive list of refinements defined by the Dublin Core Metadata Initiative (DCMI), available on the DCMI site.	Optional		
Coverage	The extent or scope of the content of the resource. Usually refers to geographic coverage or a time period (temporal).	Optional	Thesaurus of Geographic Names	
Rights	Information about rights held in and over the resource. This should be a short and simple statement defining terms of use for the digitized resource. (Mandatory for Collaborative Digitization Program)	Required		

Appendix D

CARLI Digital Collections

Guidelines for the Creation of Digital Collections

This document sets forth guidelines for digitizing materials for the CARLI Digital Collections. The issues described concern image quality, file formats, storage and access.

Image Collections

Although no universal standards for quality image capture exist and technical standards are constantly evolving, the CARLI Digital Collections will adhere to the best practices adopted by recognized leading institutions.

Digital Images

A digital image is a two-dimensional array of small square regions known as pixels. In the case of a monochrome image, the brightness of each pixel is represented by a numeric value. Gray-scale images typically contain values in the range from 0 to 255, with 0 representing black, 255 representing white and values in between representing shades of gray. A color image can be represented by a two-dimensional array of Red, Green and Blue triples, where 0 indicates that none of that primary color is present in that pixel and 255 indicates a maximum amount of that primary color.

Creating Images

At least one copy of a digital master or archival image file should be created for each object photographed or scanned. From that master file, at least two derivative files will be created:

- An access image (an image used for detailed on-screen viewing)
- A thumbnail image (for fast access during search, browse and retrieval)

A total of three types of images should be generated when an object is digitized:

Appendix D

Master Image	Access Image	Thumbnail Image
• Represents as closely as possible the information contained in the original • Uncompressed, or lossless compression • Unedited • Serves as long term source for derivative files and print reproductions • Can serve as surrogate for the original • High quality • Large file size • Stored in the TIFF file format	• Used in place of master image for general web access • Generally fits within viewing area of average monitor • Reasonable file size for fast download time; does not require a fast network connection • Acceptable quality for general research • Compressed for speed of access • Usually stored in JPEG or JPEG2000 file format	• A very small image usually presented with the bibliographic record • Designed to display quickly online; allows user to determine whether they want to view access image • Usually stored in GIF or JPEG file formats • Not always suitable for images consisting primarily of text, musical scores, etc.; user cannot tell what content is at so small a scale

from Western States Digital Standards Group, Digital Imaging Working Group, *Digital Imaging Best Practices*, http://www.cdpheritage.org/digital/scanning/documents/WSDIBP_v1.pdf, January 2003.

Master Images

Due to the stress of digitizing unique materials, a digital master should be generated for every object created. The digital master image represents as accurately as possible the visual information in the original object. This image's primary function is to serve as a long term archival record, as well as a source for derivative files and printed materials. Digital master files are measured in ppi (pixels per inch). Master files are most often saved to a designated server or other long-term storage device (such as CD-Rs).

Master images should be scanned at an appropriate level of quality to avoid re-handling of any original materials. Scanned master images should not be edited for any specific output or use, and should be saved as large TIFF files with lossless or~~,~~ no compression.

Appendix D

Creating digital master files:

- Guidelines for file size and resolution of digital master files will vary by collection based on end user needs, sizes and types of original objects, software specifications, available file storage space, etc.

- Each library should develop specific scanning guidelines based on individual collection needs and requirements.

- Where possible, scanning guidelines for creation of digital master files should follow the specifications outlined in the CDL Guidelines for Digital Images: Guidelines for Digital Master Files: http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#guidelinesmaster

Derivative Images

Derivative files are used for editing and enhancement, conversion to different formats, and presentation or transmission over networks. For each master image, two derivative files are created: an access image (for more detailed onscreen viewing) and a thumbnail image (for searching and browsing). In the case of collections using CONTENTdm, the software can be configured to automatically generate access and thumbnail images from the master file.

General Guidelines for Creation of Derivative Files:

	File Format	Pixel Array and Resolution	Bit Depth
Access Image	JPEG or JPEG2000	1024-3000 pixels across the long dimension (72 – 300 ppi)	8 bit grayscale or 24 bit color
Thumbnail Image	GIF or JPEG	100-200 pixels across the long dimension (72 ppi)	4 - 8 bit grayscale or 8 – 24 bit color

from CDL Guidelines for Digital Images: Guidelines for Derivative Files, http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#guidelinesderiv, March 10, 2005

Appendix D

File Naming Conventions

Each digital object in a collection should be assigned a unique identifier. Unique Identifiers should follow a consistent naming format to ensure ongoing identification and retrieval of digital files.

Guidelines for file names will vary by collection based on local needs and specifications. Each library should develop specific file naming conventions based on individual collection needs and local requirements.

Monitor Calibration

Monitors used for image editing and color correction should be calibrated according to the following specifications:

- Set to 24 millions of colors
- Set monitor Gamma at 2.2
- Color temperature at 6500 degrees K

Monitor calibration software can be selected and purchased by member libraries and will vary depending on local budgets, equipment and software specifications.

Appendix D

Text Collections

Text materials include printed matter, photocopies, typed or laser printed documents, may include some line drawings, graphic illustrations, manuscripts, music scores, blueprints and plans.

When scanning text documents, spatial resolutions should be based on the size of text included in the document and resolutions should be adjusted accordingly. Documents with smaller printed text may require higher resolutions and bit depths than documents that use large typefaces.

The following chart specifies basic guidelines for text document capture:

	File Format	Pixel Array and Resolution	Bit depth
Master Image	TIFF	4000-6000 pixels across the long dimension. Adjust the scan resolution to produce a Quality Index (QI) measurement of 8 for the smallest significant character. For more information about QI, see the NARA guidelines.	1-bit bitonal mode, 8-bit grayscale, or 24-bit color
Access Image	JPEG or JPEG2000	1024-3000 pixels across the long dimension (72 – 200 ppi)	1-bit bitonal or 8-bit grayscale: 72-200 dpi

based on: CDL Guidelines for Digital Images, http://www.cdlib.org/inside/diglib/guidelines/bpgimages/, June 7, 2005.
NARA Guidelines: http://www.archives.gov/research/arc/digitizing-archival-materials.pdf

Machine Readable Text

Machine readable text results either from a scanning and conversion process performed on textual materials or from manually transcribing text with a word processor.

In digital library collections, text files are often stored in such a way that they can be displayed on-screen, and they are often processed and indexed so that the content is searchable. Many options exist for digitizing and indexing text. Among them are:

- **Optical Character Recognition**
  OCR is a system that reads text and translates the image into a form the computer can manipulate. The process transforms

a bitmapped image of printed text into text code, thereby making it machine readable.

- **Transcriptions**
  Text that is difficult to read, especially handwritten manuscripts should be considered for transcription. Transcribed text, particularly if it is encoded with markup languages, helps the researcher navigate and search long documents. Transcription presents its own problems – it can be labor intensive and cost prohibitive.

- **Character Encoding**
  Character encoding is the assignment of a computer code to each of the letters in the document. A text encoded with a markup language provides searchability. Recognized text in access copies may be delivered in a variety of text formats, including HTML, ASCII, XML in EAD, TEI or other accepted standard depending on the needs of the project. Participants in the American Memory Project at the Library of Congress, use SGML in a DTD (Document Type Definition) based on the TEI (Text Encoding Initiative) Guidelines. Since SGML viewers are not yet freely available for viewing SGML over the Internet, an HTML version can be derived from the SGML version for widespread viewing online.

Text based materials in the CARLI Digital Collections may be handled in various ways. Methods will depend on factors such as library resources, quality of the original materials, software requirements, and end user needs.

Appendix E

MINIMUM GUIDELINES FOR DIGITAL IMAGE CREATION

ORIGINAL MATERIAL	DIGITAL MASTER			SCREEN DISPLAY			THUMBNAIL DISPLAY IF NOT AUTO-GENERATED IN DOMS		
	Pixel array	Bit depth	File format	Pixel array	Bit depth	File format	Pixel array	Bit depth	File format
Text document	4000-6000 across the long dimension	Bitonal, 8 bit grayscale, or 24 bit color	TIFF	Minimum 1024 across the long dimension	Bitonal, 8 bit grayscale, or 24 bit color	JPG or JP2	100-200 across the long dimension	Bitonal, 4-8 bit grayscale, or 8-24 bit color	GIF or JPEG
Illustrations, Maps, Manuscripts, Mixed Formats, etc.	4000-6000 across the long dimension	8 bit grayscale or 24-48 bit color	TIFF	Minimum 1024 across the long dimension	8 bit grayscale or 24 bit color	JPG or JP2	100-200 across the long dimension	4-8 bit grayscale or 8-24 bit color	GIF or JPEG
Film, slides & negatives: 35 mm and medium format up to 4x5 in.	4000-6000 across the long dimension (excluding mounts and borders)	8 bit grayscale or 24-48 bit color	TIFF	Minimum 1024 across the long dimension	8 bit grayscale or 24 bit color	JPG or JP2	100-200 across the long dimension	4-8 bit grayscale or 8-24 bit color	GIF or JPEG
Photographic Materials: 8X10 in. or smaller	4000 across the long dimension	8 bit grayscale or 24-48 bit color	TIFF	Minimum 1024 across the long dimension	8 bit grayscale or 24 bit color	JPG or JP2	100-200 across the long dimension	4-8 bit grayscale or 8-24 bit color	GIF or JPEG
Photographic Materials: Equal to or larger than 8x10 in. up to 11x14 in.	6000 across the long dimension	8 bit grayscale or 24-48 bit color	TIFF	Minimum 1024 across the long dimension	8 bit grayscale or 24 bit color	JPG or JP2	100-200 across the long dimension	4-8 bit grayscale or 8-24 bit color	GIF or JPEG

*Guidelines are based on the CDL Guidelines for Digital Images: http://www.cdlib.org/inside/diglib/guidelines/bpgimages/, June 7, 2005