Imagining the Northwest : A Digital Library Partnership in Oregon

Corey Harper, Nathan Georgitis, Carol Hixson
University of Oregon, USA
{charper, nathang, chixson}@darkwing.uoregon.edu

Abstract

This paper documents the development of a digital library of still images created by photographer Lee Moorhouse on the Umatilla Indian Reservation at the turn of the 20th century. The University of Oregon Libraries, working with the Tamastslikt Cultural Institute of the Confederated Tribes of the Umatilla, developed a Dublin Core compliant metadata structure. The metadata structure accommodates descriptive metadata from different cultural perspectives as well as technical and administrative metadata about multiple manifestations of images. The paper also details the development of scanning and color management methodologies, the design and application of controlled vocabularies, and the implementation of the metadata structure in the CONTENTdm Software Suite. The authors discuss the challenges of creating a flexible and interoperable metadata structure and explore the use of XSLT as a mechanism for transforming metadata exported from CONTENTdm into more granular formats. The paper concludes with a discussion of recommended best practices and future directions. The project is currently in progress and will be completed in Fall 2003.

Keywords: *Digital Libraries, North American Indians, Photography – Databases, Dublin Core, Metadata Interoperability.*

1. Introduction

In the fall of 2000, the University of Oregon (UO) Libraries launched the Digital Library Initiative (DLI) as part of an ongoing effort to improve its resources and services. A DLI working group recommended the purchase of a mass storage unit (MSU) and the CONTENTdm Software Suite (http://contentdm.com/), and prepared a policy governing the use of the use of the MSU and a statement of best practices for access to digital collections. The MSU was assembled locally in April and May of 2001. The 300GB unit is relatively small, but the Digital Library Initiative anticipated expanding the MSU with additional units.

In the fall of 2002, the UO Libraries formed the Metadata Implementation Group (MIG) (http://libweb.uoregon.edu/catdept/meta/metahome.html) to continue the work of the Digital Library Initiative. The group included members of the Catalog Department, Special Collections and University Archives, the Document Center, the Visual Resources Center for the Architecture and Allied Arts Library, the Science Library and the Collection Development and Acquisitions Department. The Libraries charged the group with creating a digital library collection in CONTENTdm and developing a flexible metadata structure based on controlled vocabularies and existing metadata standards, including the DCMES [1].

1.1. The Collection

The Major Lee Moorhouse Collection, 1897-1920, is a cornerstone of the Libraries' holdings of rare photographs. The collection comprises over 8,000 images, most of which are glass-plate negatives created by Major Lee Moorhouse, an amateur photographer and surveyor, insurance broker, civic booster and government agent to the Umatilla Indian Reservation. The Tamastslikt Cultural Institute (TCI) owns several thousand prints of Moorhouse images, including some made from glass-plate negatives owned by the Libraries.

Image 1. Poker Jim Chief of Round Up, Pendleton, OR

The Moorhouse Collection includes images of ceremonies, events and landscapes in and around the Umatilla Indian Reservation near Pendleton, Oregon. The collection also includes portraits of Native Americans from many tribes, such as the Cayuse, Walla Walla, Umatilla, Warm Springs, and Nez Perce. The portraits include regalia, ceremonial objects, weapons, and tools, among other artifacts (Image 1). However, as was common practice at the time, Moorhouse often posed his subjects with artifacts from

other tribes and from his own collections. Consequently, it is often difficult to distinguish the cultural content of the images from the contributions and manipulations of the photographer.

1.2. The Project

With a grant from the Northwest Academic Computing Consortium (NWACC), the University of Oregon Libraries formed a partnership with the TCI and the Western Interstate Commission for Higher Education (WICHE) to create a digital library collection of selected Moorhouse images called Imagining the Northwest (http://libweb.uoregon.edu/speccoll/image_svcs/imagining/)

The purpose of the project was to make the Moorhouse images available online in a culturally balanced context for use by the TCI, the people it serves, and the general public. One of the goals of the partnership was to give the tribes the opportunity to describe their cultural record in their own words by creating image descriptions. Another goal of the project was to preserve images by transferring them from fragile glass plates to a more stable medium. The partners agreed to create high-quality analog and digital surrogates for the glass plates in the form of film negatives and digital image files in order to limit use of the plates and simultaneously to improve access to the images.

The project partnered the UO Libraries, owner of the images, with TCI, an organization uniquely suited to interpret and describe them, and WICHE, an organization committed to helping educators reach underserved communities through the use of technology. Each of the organizations contributed to the project. The TCI selected images with significant cultural content and contributed rich image descriptions. The Libraries digitized the images, created stable film negatives, developed a metadata structure to accommodate multiple descriptions, and provided additional free-text descriptions and controlled vocabulary terms to the selected images. WICHE's role was to design the project Web site and develop the user interface to the collection. Many individuals participated in the project and they are acknowledged online at: http://libweb.uoregon.edu/catdept/meta/moorhouse/people.html.

2. Development of a Metadata Structure

Baca [2] defines and categorizes different types of metadata, outlines its importance, and describes the life cycle of an information object. She notes that developers of digital information systems should consider which metadata schemas to apply in order to best meet the needs of a particular user group. They must also decide which aspects of metadata are essential and how granular each type of metadata must be. The metadata schemas being applied must also be the most current versions. She provides crosswalks between different metadata standards and links to standards and standards-setting bodies.

SEPIA [3] notes that the success of a digitization project depends on the quality of its descriptions and that collections need a reliable and standardized set of descriptive data elements to be interoperable. It provides a list of data elements for photographic collections, defining them at a high level of detail and recommending best practices for each element. It recommends standards on which to base the content of different elements and provides links to many of the standards. It also lists 21 core elements and maps them to Dublin Core. The National Initiative for a Networked Cultural Heritage [4] notes that accurate metadata for the objects in a digital collection is as important as the digital surrogates themselves. It also provides a brief appendix on a number of metadata standards, with links to more detailed information.

The Institute of Museum and Library Services, Digital Library Forum [5] specifies several metadata principles that are essential for good digital collections. Metadata should be appropriate to the materials in the collection and the users of it, support interoperability, include a clear statement on the conditions and terms of use for the digital object, support the long-term management of objects in collections, and be authoritative and verifiable.

In keeping with the principles of metadata management and the parameters of the project, the Metadata Implementation Group (MIG) developed a flexible and interoperable metadata structure to serve the needs of the collection and to meet the demands of varied users. The data dictionary for the metadata structure is available on the project Web site at the following address: http://libweb.uoregon.edu/catdept/meta/Moorhouse1.0a5-19.xml.

The MIG began by defining its general approach to the description of digital library collections. Group members endorsed the creation of a unique data dictionary for each collection and the use of a crosswalk standard, in the form of simple Dublin Core, for resource discovery both within and across collections. To reduce inconsistency in metadata application and to optimize interoperability, the group also supported the use of controlled vocabularies and encoding standards throughout the metadata structure.

The need to describe different manifestations of images, including versions optimized for browsing or printing, was debated early in the project. While acknowledging the need to track these manifestations, the group questioned the need to describe each in a separate record and decided to include core metadata for all manifestations of an image within a single metadata record. This approach allows metadata to be extracted and used to populate separate records in the future.

After reviewing the Open Digital Rights Language (ODRL), Version 1.0 (http://odrl.net/), the group decided that a comprehensive set of rights management metadata elements was beyond the scope of the project, and instead created a single Web page detailing rights issues for each collection. The University Archivist drafted a statement

outlining rights issues for all digital image collections (http://libweb.uoregon.edu/catdept/meta/moorhouse/rights.html).

The Moorhouse project team, a subset of the Metadata Implementation Group, began work on the Moorhouse data dictionary by evaluating separate databases previously created by the UO Libraries and the TCI to describe Moorhouse images. The team reviewed the descriptions in the databases and affirmed the value of the different perspectives they presented. A data dictionary was created accommodating both sets of descriptive metadata in discrete fields rather than merging the two descriptions within a single set of elements. This data dictionary maps similar metadata elements, such as UO Title and TCI Title, to common Dublin Core elements. Although this strategy increased the size and complexity of the metadata structure, it preserved the richness of the metadata associated with the images, while still providing the extensibility of the simple Dublin Core framework.

After reviewing a number of standards concerning administrative and technical metadata, the team concluded that the NISO data dictionary of technical metadata for digital still images [6], though still in draft form, is the most thorough. The project team identified the elements in the NISO dictionary that were relevant to the methods and materials of the project, many of which are common to the *TIFF 6.0 Specification* [7] and Harvard's *Administrative Metadata for Digital Still Images* [8].

These elements were incorporated into the Moorhouse data dictionary in blocks relating to image attributes, image production, and image source. In some cases the team merged metadata from multiple elements into a single field, but retained the granularity of the NISO metadata by creating structured data values with standard delimiters. The encoding standards and controlled vocabularies of these elements were retained without exception. A crosswalk will guide the export of metadata in keeping with those standards.

2.1. Metadata Capture and Project Workflow

The next step was to develop a work plan to guide the creation and description of digital images. Arranging the work of the project into seven stages—selection, image capture, image enhancement, file processing, record completion, record review, and final approval—the project team considered the skills required to complete each stage and assigned personnel accordingly. Project team members evaluated the plan of work to determine the best points at which to capture or create different types of metadata and developed procedures to guide the creation and description of the digital images at each stage. The project team created a record template in CONTENTdm with default values for standard elements, and arranged the elements in the template in the order in which they were addressed during the plan of work.

At the selection stage, a representative from the TCI chose images for inclusion in the digital library collection in consultation with the Library's Coordinator of Preservation & Digital Services. Project staff added the selected images to the work log and a student assistant retrieved the glass-plate negatives from storage. The project metadata editor collected metadata for the selected images from the UO and TCI databases, as well as a subject index believed to have been prepared during a Works Progress Administration (WPA) project in the 1930s. To facilitate data entry, the metadata editor made this metadata available to the project technicians in both print and electronic form.

At the scanning stage of the project, technicians scanned the glass-plate negatives in color in transparency mode and adjusted the tonal range of the images. These TIFF images became the master image files. Technicians captured image source metadata, such as the condition and dimensions of the negative, and general descriptive metadata, including the title on the image, the photo number, and the date of the photograph. The Image Services Supervisor trained the technicians in the handling of the negatives, the evaluation of their condition, and the use of the scanning software.

At the image enhancement stage, technicians evaluated the quality of the TIFF images and enhanced them for electronic presentation as JPEGs. Technicians captured image production metadata, including processing software and processing methodology. Technicians also selected terms from a local, controlled vocabulary to describe processing actions or image enhancements.

At the file processing stage, technicians created derivative image files, including compressed 100 dpi JPEG images for Web display. They captured relevant technical metadata concerning these manifestations. The project team identified the TIFF file header as the source of most technical metadata and used a TIFF file editor to facilitate access to file header information. At this stage, technicians also used an MD5 checksum utility to create checksum values for image files.

At the record completion stage, Catalog Department staff and librarians examined the images and completed the metadata records. They created high-level descriptive metadata, including subject description, content description, and supplied titles.

At the record review stage, Catalog Department librarians reviewed the enhanced images and the completed metadata records and released them to the TCI for review and further description. The TCI reviewed the images and records and contributed corrections and additions. Finally, project coordinators reviewed the enhanced images and the completed metadata records and approved them for publication in the digital library collection.

2.2. Scanning Methodology and Color Management

The Image Services Center, a division of the Libraries' Special Collections and University Archives, worked with the Metadata Implementation Group to develop color-management procedures for scanning glass-plate negatives. The group reviewed the available literature on digital imaging and color management, including the publications of the Digital Library Federation (DLF) [9] and the National Initiative for a Networked Cultural Heritage (NINCH) [4]. Following a trial of the equipment and workflow, the project team identified five areas for improvement, including workstation equipment, equipment calibration, viewing environment, color profiles, and capture settings.

The Image Services Center recommended the purchase of new equipment for two digital imaging workstations, including Dell computers with 17-inch, flat-screen Trinitron aperture-grille monitors. The Center evaluated a number of capture devices according to the criteria outlined in the DLF document and decided that UMAX PowerLook III flat-bed scanners with UMAX transparency hoods offered sufficient quality and versatility while remaining affordable.

The Image Services Center also recommended methods for equipment calibration. A staff member from the Center used Adobe Gamma, a control panel utility in Adobe Photoshop, to calibrate the workstation monitors in the scanning laboratory. The staff member adjusted the brightness, contrast, white point, and RGB gamma for each workstation. These calibrations were then saved as the International Color Consortium (ICC) profiles for these monitors. ICC device profiles correlate the color space of a device with a defined reference color space, which allows images to be represented accurately in the color spaces of different input, output and display devices.

Naturally, these adjustments depended on the color acuity of the staff member and the viewing environment in the laboratory. Although this method of calibration is subjective, the team decided that it was acceptable given the materials and goals of the project. An effort was made to optimize and standardize the viewing environment in the laboratory to limit inconsistencies in color perception.

The Image Services Center recommended Silverfast Ai Version 6.0 to generate color profiles for scanners through the use of standard IT8 targets and for use as a scanning utility. A staff member from the Image Services Center used the software with transparent and reflective IT8 color targets to generate ICC profiles for the scanners. All images were scanned in 48-bit RGB (16 bits per channel) and converted automatically to 24-bit RGB (8-bit per channel) when saved as TIFFs. Upon import into Adobe Photoshop, the color profiles of scanned images were converted to Adobe RGB (1998) because of its wide color gamut. The metadata for each master TIFF file includes a reference to this ICC profile. Since current Web browsers ignore embedded ICC profiles and default to sRGB, the group decided to convert image files meant for the Web to sRGB and not embed ICC profiles.

The project team acknowledged that the management of color in digital imaging projects by either objective or subjective means is bound to be imperfect. With this in mind, every effort was made to develop a consistent workflow with adequate quality control measures.

3. Controlled Vocabulary Design and Application

Over a period of several months, both the MIG and the Moorhouse project team discussed the importance of controlled vocabularies in providing subject access to digital image collections. They also reviewed some specific challenges related to the Moorhouse collection. These discussions were influenced by a review of other digital image collections, various guidelines, best practice documents, and thesauri.

Shatford [10] discusses principles of providing subject access to pictorial materials, noting the different perspectives of various user groups and that the same user will approach the same picture differently at different times. Layne provides guidance in classifying the subjects of a picture by noting that its different facets may be defined initially as answering the questions of Who? What? When? and Where? She then further subdivides these facets based on the aspects Of and About. She advises catalogers to bear in mind the nature and the intended audience of a collection of images.

Will [11] outlines the need for building a thesaurus as a means of documenting rules and decisions made over time and maintaining consistency in the application of terms. He notes that a simple list of names without some rules for application will quickly become problematic and lays out three rules which should guide every thesaurus: 1) use a limited list of indexing terms, but plenty of entry terms; 2) structure terms of the same type into hierarchies; and 3) remind users of other terms to consider.

SEPIA [3] provides a list of data elements, defining them at a great level of detail and recommending best practices for each element. In the section on descriptors and subject headings, it recommends the development of local lists, tightly controlled by sticking to firm rules.

The *Thesaurus for Graphical Materials I: Subject Terms TGM 1* (http://www.loc.gov/rr/print/tgm1/) gives information about the building of the thesaurus and provides a controlled vocabulary for describing still images. The fact that it draws on terms from other established thesauri illustrates the flexible approach needed for building a controlled vocabulary for image collections.

In early discussions about controlled vocabularies for digital collections, MIG members acknowledged that image description would be carried out by a variety of staff, from students to collection curators. There often would not be time to provide elaborate training or expert review of cataloging. In addition to being guided by accepted principles regarding the building of a controlled vocabulary for digital collections, they also agreed to the following

local principles: 1) As much as possible, build the list of controlled terms beforehand. When no appropriate term is available to indexers from a controlled list, allow indexers to input an appropriate term and to flag the record for later review. This policy derives from the *TGM1* principle that terms are added only as topics are encountered in the course of cataloging; 2) Use multiple controlled lists to cover different aspects of the subject, if needed; 3) Use broader terms to describe the subject if more specific terms are not available; and 4) Provide any hierarchal relationship among terms by means of a search interface. These principles were followed in the execution of the Moorhouse project.

3.1. Designing the Vocabularies for Moorhouse

Prior to the grant-funded partnership with the UO Libraries and WICHE, TCI had described nearly 900 of their Moorhouse images in a local database using PastPerfect software (http://www.museumsoftware.com/). PastPerfect comes equipped with *The Revised Nomenclature for Museum Cataloging: A Revised and Expanded Version of Robert G. Chenhall's System for Classifying Man-Made Objects*. Using the *Nomenclature* system as a foundation, the software organizes each object name into 11 categories and 100 sub-categories. It also provides basic authority control, checking new object names against the approved list as data is entered and allows for the revision of and addition to the controlled vocabulary to suit local needs. To supplement the list of terms supplied in the software, TCI staff also made use of the *Sears List of Subject Headings*, 17th edition. In the process, they developed a unique controlled vocabulary, including the concept of overarching classes and subject terms, which they had applied with a fairly high level of consistency.

Also prior to the grant, the Libraries' Special Collections and University Archives staff had created a Microsoft Access database in which they recorded descriptions and subject terms for previously digitized images from the Moorhouse collection. Most of the digitized images had been created in response to patron requests for reproductions of a particular image. Because they developed the database to provide a rudimentary tracking and searching mechanism for library staff, rather than to provide public access to the collection, the subject terms they applied were not selected from a controlled vocabulary and were not applied with great consistency.

Many of the prints that TCI staff worked with were annotated with information about the people and places depicted in them. As TCI staff described the images, they attempted to correct misidentifications of people depicted in the photographs. As they systematically reviewed and described the images, they also observed that the same piece of tribal regalia appeared repeatedly in different images with different people, sometimes in obviously staged settings. They determined that it was important to identify each piece of regalia uniquely and to track the regalia across images. In effect, they decided to apply authority control to specific pieces of regalia. The project team decided to retain the information about regalia in a separate field in the database.

One aspect that distinguished this project from others was the desire to make it possible to provide different descriptions or interpretations of the same image from different cultural perspectives. TCI staff had made it clear that people from different tribal backgrounds might have substantially different interpretations of the content of the same image; the same person or ceremony being depicted in an image might be known by different names to different groups. From the start, project members sought to provide for this multi-dimensional description of image content in the design of the database.

As recommended by many of the sources consulted, the group examined other collections of digital still images, finding the Edward S. Curtis collection of the North American Indian (an American Memory project between the Library of Congress and Northwestern University Library) particularly useful (http://memory.loc.gov/ammem/award98/ienhtml/curthome.html). The way that the Curtis collection provided groupings for subject terms seemed to parallel the TCI use of classes. The group determined that such an approach could be accommodated within the context of CONTENTdm via the creation of Web pages with pre-defined searches.

The four members of the Catalog Department on the Moorhouse project group began to develop a controlled vocabulary for the grant project. Group members searched various indexes or thesauri for sample terms that had been used previously to index the Moorhouse collection. For topical subjects, they searched the *Art and Architecture Thesaurus (AAT)*, the *Library of Congress Subject Headings* (*LCSH*), *TGM 1*, and *Thesaurus for Graphic Materials: Genre and Physical Characteristic Terms (TGM2)*. Following the sample searches, project members decided that *TGM 1* and *LCSH* were more likely to contain terms and cross-references useful in describing the content of these photographs; *AAT* seemed to lack appropriate terms for many of the concepts depicted in the collection.

One of the first steps taken was to merge the subject and descriptive data from the existing TCI database and the UO database. Project staff were able to merge the descriptive data because the same image numbers had been used in both databases. The data combined from the TCI spreadsheet and the UO database proved to be an excellent foundation for the construction of a local controlled vocabulary. The finalized workflow involved searching the list of combined subject terms only in *TGM 1* and *LCSH* in order to establish a single term with cross references for each concept. Staff members from the Catalog Department searched the combined, de-duped list of subject terms in *TGM 1* and recorded any matching terms and cross references found. If the search of *TGM 1* failed to return a match, staff repeated the process using *LCSH*. The

resulting controlled vocabulary list was created in the text format that CONTENTdm accepts and was loaded into the database as the basis of the controlled vocabulary. The sources of each term and cross-reference are tracked in an additional collection-level document. This process is illustrated in Figure 1 below.

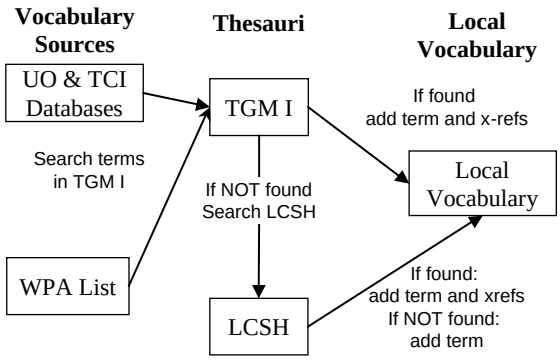

Figure 1: Local controlled vocabulary design

The group discussed the possibility of using Encoding Scheme Qualifiers to define separate Dublin Core (DC) Subject fields for each vocabulary. There was some concern that using such a detailed application of DC would make it more difficult for non-experts to apply the terms. Of more practical concern was that such an approach did not appear to be a viable option within the CONTENTdm framework. CONTENTdm only provides mapping to Qualified Dublin Core for Element Refinement Qualifiers, not for Encoding Scheme Qualifiers. The group briefly discussed providing the source of a term as an annotation, such as Umatilla Indians (LCSH). However, in order to avoid cluttering up the search and display, the group decided to document the source of a particular term at the collection level, rather than as part of the term itself as applied to single images.

In addition to providing a field for locally-controlled subject terms, the project group retained the TCI terms in their original form in two separate fields (one for subject terms and another for classes) and sought a way to utilize the broader classes effectively in indexing and in a search interface. Only TCI staff were authorized to assign the terms in these fields.

The four project leaders from the Catalog Department were reluctant to discard the rich subject indexing provided by the WPA list. The list contained detailed information about people, places, and events depicted in the images that would be invaluable to the indexers in choosing from and adding to the locally-developed list of terms. The project leaders converted the WPA list to electronic form, sorted it by image number, and made it available to indexers. After that step was completed, they introduced the volunteer indexers from the Catalog Department to all the locally-developed tools and trained them in some of the principles of applying subject terms to image materials. TCI staff would subsequently review all UO-originated subject analysis and suggest corrections and clarifications.

4. Creating a Collection in CONTENTdm

CONTENTdm was developed at the University of Washington as a management system for digital images. The software platform is built on open standards, including the Dublin Core (DC), Visual Resource Association (VRA) Core, and the eXtensible Markup Language (XML). XML is included as an option for exporting metadata for use in other systems. The most recent release includes an interface to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which also utilizes XML to communicate with OAI harvesters. The CONTENTdm implementation of OAI supports interaction with any and all of the collections defined in CONTENTdm through a robust cross section of OAI verbs. All metadata is returned in XML documents conforming to the oai_dc format, although the structure of these XML documents differs from the structure of the XML that can be exported directly through the Collection Administration interface.

Preliminary efforts to test the CONTENTdm system enabled members of the Metadata Implementation Group to work with its functionality for batch import of image files and corresponding metadata from a comma-delimited text file. CONTENTdm effectively processes text files if commas consistently separate database elements and line breaks separate individual records. Prior to extracting the text file, project team members updated the database so that unique image identification numbers matched the image file names, in conformance with the input requirements of CONTENTdm. This correlation enables the system to map each 'row' in the import file to a specific image, while also allowing the user to choose how each 'column' maps to the metadata fields that are defined in the collection. Alternately, implementers can choose to set default values from a template, although the template cannot be used in conjunction with batch import of images and text from a database. This approach was taken in the production database due to the necessity of having a group of staff apply metadata at the time of image import, and because of the relative sparseness of the original descriptive metadata.

4.1. Defining a Data Dictionary in CONTENTdm

The Dublin Core and VRA Core are provided as potential sources of field properties at the time of collection creation in CONTENTdm. There are a number of options for defining metadata field properties that can be revised even after the database for a collection has been populated. For the Moorhouse project, the DCMES is used as the default template. The following table illustrates the customization options for element attributes. CONTENTdm also provides control over the display position for each element.

Table 1. Customizable Attributes for CONTENTdm Metadata Elements

Attribute Name	Customization Options
Field Name	Editable Free Text
Dublin Core Mapping	Drop down menu provides support for both Simple and Qualified Dublin Core
Data Type	Text, Date or Full Text Search
Large Field	Yes or No
Searchable	Yes or No
Hidden	Yes or No
Controlled Vocabulary	Pages for enabling and administering

4.2. Difficulties managing metadata with CONTENTdm

The options discussed in the preceding section can be altered extensively as collection development continues and the requirements of an individual collection become clear. However, there are some limitations inherent in the manner in which CONTENTdm implements the Dublin Core and XML specifications.

There are three specific shortcomings with CONTENTdm's implementation of DC, which have an effect on the implementation of XML. The first of these relates to treatment of the 'dumb down' principle, the second relates to element repeatability, and the third concerns the Dublin Core premise that all elements are optional.

The 'dumb down' principle is one of the basic premises of Dublin Core. The idea is that the element set can be qualified and enhanced, but that applications that accommodate Simple DC can still interpret the basic meaning of qualified versions of elements. This concept is the lynchpin of the extensibility and interoperability of the DCMES. In CONTENTdm, the dumb down principle is only enforced when searching across multiple collections. Multiple fields mapping to subject (e.g. one using TGM, one free text keywords) cannot be searched as a unit within an individual collection; one can only search one field at a time or across all indexed fields. Cross collection searching uses Dublin Core fields and searches any field that is mapped to the appropriate DC element.

Second, elements are not repeatable in the Dublin Core sense of the word. The only way to repeat elements is to enter a free-text value for an element in instance metadata, and including a delimiter to separate occurrences. This same method is also used for elements that derive their contents from a controlled vocabulary with delimiters, in the form of semi-colons, which are provided automatically. This does allow for multiple values for an element, but upon exporting the metadata, the resulting XML lumps all values in one tag set, including the delimiters that mark individual occurrences.

This repeatability issue becomes even more muddled when looking at metadata exported for instances in which multiple elements in the local application profile map to the same Dublin Core element. In addition to repeating the tags for separate occurrences of the same element, the resulting XML also concatenates separate elements that map to the same DC element into a single XML tag for that element. This presents an obstacle to producing XML instance metadata that conforms to a schema derived from the local data dictionary or application profile. Additionally, it is difficult to translate the resulting XML into other formats, which poses a deterrent to customizing the implementation of OAI. A solution to these problems using XSLT is discussed in Section 5.

Finally, there is a problem with CONTENTdm's implementation of the Dublin Core premise that all fields are optional. This last issue was actually fairly easy to remedy but, depending on how far along in development a project is, the solution can be somewhat time consuming.

When the initial Dublin Core field properties are defined for a collection, CONTENTdm mandates a Title element. The contents of this field display as index titles for thumbnails when a collection is browsed. The MIG had discussed at length the relative value of titles for untitled photograph collections. Some members of the group felt that cataloger supplied titles were misleading, and that a title field should only be populated if the title appeared on the glass-plate negative itself. An effective compromise was reached by developing five separate variants on the title field, and only populating the ones that are germane to a specific image. The five instances of Title are implemented in the application profile using locally defined qualifiers. None of these title fields are mandatory.

The only mandatory field in the data dictionary for this collection is the unique alphanumeric photograph number, which has been mapped to DC Identifier. It had already been determined that this identifier field should serve as the labels for thumbnail images when browsing, so it was decided to make this the mandatory element. Rearranging the list so that the identifier element appeared first did not affect the requirement that a title be provided. A solution was found by renaming the 'Title on Object' field 'Photo Number', and changing its mapping from DC Title to DC Identifier. Likewise, the Photo Number field became Title on Object, and the values of these two fields were swapped.

5. Interoperability and exporting metadata

There are two distinct methods for generating XML of the instance metadata for any given collection in CONTENTdm. One method uses the export metadata functionality provided by the collection administration interface. Alternately, or additionally, the set of CONTENTdm collections can be made available as an OAI data provider, and standard OAI queries can be used to extract XML instance metadata, other information about the collections and descriptions of the system as a whole.

Unfortunately, the XML formats exported by the system through the Web interface and through the OAI interface are not interoperable with one another.

As mentioned earlier, CONTENTdm has the functionality to export collection-level instance metadata in XML. The implementation of this feature is at odds with *Guidelines for implementing Dublin Core in XML* recommendation [12]. Recommendation 5 of this document states, "multiple *property values* should be encoded by repeating the XML element for that *property*." CONTENTdm takes the opposite approach, concatenating each value for a property within the opening and closing tags of a single occurrence of that property's XML tag. Additionally, if multiple properties are mapped to a single Dublin Core element in a local project these values are concatenated and expressed in a single tag set.

5.1. XSLT and Implementing Dublin Core in XML

Fortunately, the XML that CONTENTdm produces is valid and can be processed to conform to a variety of formats using XSLT. The inclusion of HTML line breaks *
* at the end of each field enables the separation of various instances of an element. Five different fields in the Moorhouse data map to DC Subject. However, when encoded according to the local application profile, only one of these is truly unqualified DC Subject. The other four are mapped to locally defined qualifiers that refine the DC Subject element, in line with the practice recommended by Heery and Patel [13]. In the context of CONTENTdm's system generated XML the subject terms found between the second and third sets of HTML line breaks (
) are the subject terms that actually map to the unqualified DC Subject element. If the data dictionary is developed carefully and fields are populated consistently, XSLT provides an easy mechanism for parsing out and processing the fields accordingly.

Experimentation with XSLT in the context of the Moorhouse collection has yielded a set of processing instructions that transform the exported XML into formats corresponding to a local application profile and adhering to the guidelines for implementing DC in XML. The two tables below are fragments of XML and XSLT documents. Each table is divided into three rows. The first row is a fragment of instance XML exported from the CONTENTdm system, either through the default metadata export interface, or through the OAI implementation. The second row is a fragment of the XSLT that processes the given XML fragment, and the third row is the XML that the processing instructions produce. Each of these fragments has been removed from its source context, which includes namespace declarations, parent nodes, and in the case of the XSLT documents, additional processing instructions.

Table 2. Formatting data according to DC Guidelines

```
<dc:subject>TCI Terms &lt;br&gt;TCI Classes &lt;br&gt;
    Indians; Camps; Tipis &lt;br&gt; Names &lt;br&gt;
</dc:subject>
```

```
<xsl:for-each select="saxon:tokenize(string(substring-
    before(substring-after(substring-after(dc:subject,
    '&lt;br&gt;'),'&lt;br&gt;'), '&lt;br&gt;')), ';')">
    <xsl:element name="dc:subject">
        <xsl:value-of select="(.)"/>
    </xsl:element>
</xsl:for-each>
```

```
<dc:subject>Indians</dc:subject>
<dc:subject>Camps</dc:subject>
<dc:subject>Tipis</dc:subject>
```

Table 2 illustrates this concept using an XML fragment for <dc:subject> exported through the collection administration interface. The XSLT processing instruction uses nested XPATH 'substring' expressions to locate the third set of subject terms, which map to the DC Subject field.

Given the alternative XML formatting returned by the OAI interface, the XSLT extensions in the Saxon XSLT processor can produce XML that conforms to the *Guidelines for implementing Dublin Core in XML* [12]. The resulting XML adheres to a locally defined application profile combining DC elements with elements and qualifiers from a local schema. The fragment of XSLT shown in Table 3 processes the oai_dc XML and produces the same XML format that is produced in Table 2.

Table 3. XML and XSLT to process OAI data

```
<dc:subject>TCI Terms &lt;br&gt;</dc:subject>
<dc:subject>TCI Classes &lt;br&gt;</dc:subject>
<dc:subject>Indians; Camps; Tipis;
&lt;br&gt;</dc:subject>
 <dc:subject>Names &lt;br&gt;</dc:subject>
```

```
<xsl:for-each
select="saxon:tokenize(dc:subject[position()=3])">
    <xsl:if test="position()!=last()">
        <xsl:element name="dc:subject">
                <xsl:value-of select="(.)"/>
        </xsl:element>
    </xsl:if>
</xsl:for-each>
```

```
<dc:subject>Indians</dc:subject>
<dc:subject>Camps</dc:subject>
<dc:subject>Tipis</dc:subject>
```

Line breaks are not needed when processing OAI data since the XML already has a separate occurrence of the subject element for each locally defined element that is mapped to subject. In this case, the same parsing can be done using the 'position' function rather than the nested substring functions. All that remains is to ignore the last node in the tokenized string, which would contain the line break.

Variations on the two processing instructions demonstrated above, in conjunction with consistent application of delimiters during data entry, allow for parsing CONTENTdm's XML formats into a variety of other formats, greatly increasing interoperability with other XML based metadata applications.

6. Conclusion

Although no formal evaluation has been done, the Metadata Implementation Group (MIG) concluded that this metadata structure will successfully accommodate different descriptions of Moorhouse images by the UO Libraries and the TCI. The metadata enriches the collection by providing culturally specific descriptions of images, including accurate names for artifacts, individuals and places. During this process, the MIG and the Moorhouse project team developed a set of best practices that should be followed as digital library development continues at the UO Libraries. This project also helped to inform the future direction that will be taken in regard to the Moorhouse collection, as well as concerning interoperability between collections.

6.1. Recommended Best Practices

Until recently, efforts to provide access to non-textual collections were almost exclusively handled by specialists or collection curators who were familiar with the issues and the available tools. As libraries strive to make more of their non-traditional collections available digitally, the responsibility for providing access to them is being expanded beyond the specialists. Working to develop the Moorhouse collection of digital images at the University of Oregon Libraries has led to a heightened awareness throughout the institution of the unique demands of providing subject access to historical image collections.

The following guidelines, while somewhat simplistic, can be considered recommended best practices for approaching subject access to digital collections. 1) Consider the target audience and its likely approach to the collection; 2) Look at other similar collections to see how subject access has been handled; 3) Review existing guidelines or standards relevant to the type of collection (image, textual, audio, discipline-specific, etc.); 4) Take a sample of possible indexing terms and search them in a select number of relevant thesauri to determine the best source of useful terms; 5) Even if free-text descriptions form part of the strategy for providing subject access, build a local controlled vocabulary, consulting existing thesauri and documenting from where terms are taken; 6) Use terms from existing thesauri whenever practical; 7) Build in multiple entry terms (cross references) to your controlled vocabulary; 8) Consider accommodating broader, narrower, and related terms either through the underlying vocabulary structure or through a search interface; and 9) Document all decisions and the thinking that led to them.

When developing a data dictionary for use with CONTENTdm, project implementers should select an element to be mandatory before development of the collection proceeds too far. This will ensure that only this element will be treated as mandatory by the system. Additionally elements can be made mandatory through the usage guidelines described in the data dictionary. Criteria for selecting this element include ensuring that it is not repeatable, will be present for each object, and will provide a meaningful caption for browsing thumbnails.

At the time of data entry, UO's Catalog Department recommends always using a line break tag to mark the end of a field in cases where multiple fields map to the same Dublin Core element. Semi-colons separate individual terms or phrases within these fields. If this practice is followed, XSLT processing instructions can be created that work within the context of a given data dictionary to identify which portions of an element's value string map to more granular and specific locally defined elements.

6.2. Future Work

Future work on the Moorhouse collection and subsequent digital collections will expand on what has been learned and accomplished to date.

Color management is an important part of any digital imaging project and integral to color management is the viewing environment in the digital imaging lab. Although efforts were made to limit natural and fluorescent lighting, the viewing environment in the lab is not ideal and needs to be addressed in detail before the Libraries embark on another digital imaging project.

The MIG will consider how the metadata structure for the digital library collection could be expanded to accommodate descriptions of images by different tribes on the Umatilla Indian Reservation. However, the management of multiple image descriptions from different sources would require a more sophisticated use of controlled vocabularies and metadata.

In the context of the NWAC grant, the success of this project will be determined based on the use TCI makes of the collection, and by the feedback they provide. Continued communication with the Institute will show the extent to which the availability of these images contributes to the preservation of the Confederated Tribes of the Umatilla's cultural heritage. Additionally, as the collection grows to include images beyond those selected by the TCI, it will be extremely important to track what use the general population and the research community makes of Moorhouse's photographs. This need will be met by the development of a system for reviewing and analyzing the access logs generated by CONTENTdm.

As the number of digital collections developed by the libraries increases, the issue of interoperability becomes increasingly significant. A measure of success will be the ability of the Moorhouse data dictionary to serve as the foundation for other digital library projects. Ideally this

effort will also lead to the development of a core set of metadata elements for any digital collection on campus.

As these efforts continue, further experimentation with XSLT, in conjunction with a more complete implementation that supports additional metadata formats, will help create an environment where seamless cross collection and cross database searching is available to library patrons, where collections external to the University of Oregon are equally accessible, and where the University's collections can easily become part of a truly global network of digitized materials.

References

[1] Dublin Core Metadata Initiative. 2003. Dublin Core metadata element set, Version 1.1: reference description. Available at: http://dublincore.org/documents/dces/

[2] Baca, M., ed. 1998. Introduction to metadata: pathways to digital information. Los Angeles: Getty Information Institute. Available at: http://www.getty.edu/research/institute/standards/intrometadata/index.html

[3] Safeguarding European Photographic Images for Access (SEPIA), Working Group on Descriptive Models for Photographic Collections. 2003. SEPIADES advisory report on cataloguing photographic collections. Available at: http://www.knaw.nl/ecpa/sepia/workinggroups/wp5/advisory30.pdf

[4] National Initiative for a Networked Cultural Heritage. 2002. The NINCH guide to good practice in the digital representation and management of cultural heritage materials. Washington, D.C.: NINCH. Available at: http://www.nyu.edu/its/humanities/ninchguide/

[5] Institute of Museum and Library Services, Digital Library Forum. 2001. Framework of guidance for building good digital collections. Washington, D.C.: IMLS. Available at: http://www.imls.gov/pubs/forumframework.htm

[6] NISO. 2002-2003. Data dictionary: technical metadata for digital still images. Draft standard. Bethesda, Md.: NISO. Available at: http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

[7] International Telecommunication Union. 2002.TIFF 6.0 Specifications. Geneva: ITU. Available at: http://www.itu.int/itudoc/itu-t/com16/tiff-fx/docs/tiff6.pdf

[8] Harvard University Library, Library Digital Initiative. 2000. Administrative metadata for digital still images. Boston: Harvard University. Available at: http://hul.harvard.edu/ldi/resources/ImageMetadata_v2.pdf

[9] Digital Library Federation, Research Libraries Group. 2000. Guides to quality in visual resource imaging. Washington. D.C.: Council on Library and Information Resources. Available at: http://www.rlg.org/visguides/

[10] Shatford, S. 1986. Analyzing the subject of a picture: a theoretical approach. Cataloging & Classification Quarterly, 6: 39-61.

[11] Will, L. 1998. Thesaurus principles and practice. Available at: http://www.willpower.demon.co.uk/thesprin.htm

[12] Powell, A., Johnston, P. 2003. Guidelines for implementing Dublin Core in XML. Available at: http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/

[13] Heery, R., Patel, M. Application profiles: mixing and matching metadata schemas. *Ariadne.* 25. September 2000. Available at: http://www.ariadne.ac.uk/issue25/app-profiles/