Preserving Regional Historical and Cultural Heritages: A Case Study in Building Local Digital Image Libraries

Yan Zhang, Iris Godwin

School of Information Science, the University of Tennessee, Knoxville, 449 Communications Bldg. 1345 Circle Park Drive, Knoxville, TN 37996, USA
{yzhang, igodwin}@utk.edu

Abstract. Digital libraries are structured collections that can offer intellectual access and ensure the persistence over time of collections of digital works. They provide a good approach to collect, organize, disseminate and preserve regional historical and cultural materials. In the designing of a digital library, considerations of value, volume and format of the collection are necessary in producing a quality compilation. Users' needs and their information seeking behaviors must also be considered. This Derris Digital Image Library Prototype Project supported by the University of Tennessee Libraries provides a case study in building a digital image library based upon a slides collection for preserving and disseminating history and culture of the Great Smoky Mountains region of East Tennessee. The related issues of digitizing slides, formulating metadata scheme, ensuring usability, interoperability and establishing authority control of the access points and subject heading are discussed in this article.

1 Introduction

It is estimated that there are over 4,000 Gigabytes of data on the World Wide Web. This number continues to grow at an exponential rate and duplicates every 6 months [1]. With this exceptional expansion, the Internet is permeating into every aspect of daily life and becoming the main information source. People are relying on it more to look for information. However, there are some problems associated with the phenomenon of information abundance. One problem is the unstructured nature of the information on the Internet. In many cases, retrieval of information from a search includes a large amount of results that are not relevant to the query statements. Another major problem is the quality of the information on the Web. Unlike the traditional libraries, where subject experts develop collections, the Internet is an open environment with extreme freedoms; there are no central agencies or mechanisms to control the quality of the information on the Internet.

Digital Libraries may be a solution to the orderless situation of the Internet. In the early stage of its development, there was much conceptual confusion surrounding the phrase "Digital Library". The average information seekers usually regarded the Internet as the largest Digital Library in the world. With the advance of researches, most researchers have come to agree that Digital Libraries should encompass the whole information life cycle [2], including the basic functions of collecting and

organizing information sources, providing intellectual access and preserving digital objects. In 1999, the Digital Library Federation defined digital libraries as "organizations that provide the resources, including specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities" [3].

With the ability to offer quick and economic access to collections and ensure persistence over time of the digital works, the Digital Library is a good vehicle for collecting, organizing, preserving and disseminating historical and cultural materials. These materials could be in the format of print, manuscript, stele or other physical objects that are often fragile, brittle and vulnerable to environmental changes. Building digital libraries to preserve the precious heritage is meaningful for making "knowledge of the ancient traditions, the experience of change and the living reality" [4] available and accessible.

This paper reports on a project that applied existing metadata formats to a university library's special collection of slides about views, history and culture of the Great Smoky Mountains. The government founded the mountain as a national park for preserving the natural heritage in 1934. The slide collection is a donation from Mr. William Derris, an amateur photographer and owner of the Derris Motel (later the Laughing Horse Inn) in Townsend, Tennessee. Throughout his ownership of the Inn, Mr. Derris photographed guests, seasonal landscapes, flora, wildlife, personal pets, and natural wonders of East Tennessee. The Derris collection is composed of approximately 4400 slides divided into 85 series by geographic places from the 1940s through the 1960s of the Great Smoky Mountains region of Tennessee. The University of Tennessee owns the copyright of the collection. One hundred sample slides drawn from 17 series of the collection serve as a prototype to test digitization and metadata treatment.

2 Digitization

Digitization is the process of taking traditional library materials and converting them to electronic form where they can be stored and manipulated by a computer. Generally, when building a digital imaging library, one of the first things to consider is whether it is needed to digitize existing documents. Digitizing a large collection in house could be an extremely time-consuming and expensive process. Any significant image digitization projects will normally be outsourced [5]. Either way, it will be necessary to have carefully planned digitization and quality control standards. Decisions about the technical infrastructure are usually based upon the considerations of document attributes, the understanding of users' needs, and the assessment of long-term preservation plans [6].

Because the size of the prototype for Derris Digital Image Library is reasonably small, we decided to conduct the digitization on our own. Furthermore, the in-house digitization allowed us to define requirements incrementally rather than up front and impose direct control over the entire range of imaging functions. The whole

digitization process may be divided into three major components: image creation, file management and image delivery.

2.1 Image Creation

Image creation deals with the initial conversion of the slides into digital pictures with a scanner and related software. In this process, we chose the computer facilities including hardware and software, and decided image quality and format. With the help of the university library's studio, we performed several trials and finally decided that we would use the Umax Powerlook 2100XL scanner and Wintel PC to do the scanning. The Umax Powerlook 2100XL scanner has a 35mm slide tray that allowed us to scan 32 slides at one time, which saved us a lot of time. The software used in the project included WindowsXP Operating System, Magic Scan32 Version 4.4, and Adobe Photoshop 7.0.

There are three image types, master, reference, and thumbnail, needed in the project. TIFF files, created through a process of scanning, were stored as masters and used to derive reference and thumbnail files for web-delivery. The master files were not compressed. TIFF was chosen as the master file format because it is designed as an industry standard for image file exchange; it is highly flexible and platform independent [7]. In addition to file format, resolution is a factor that affects the image quality. Higher resolution allows for more detail and subtle color transitions in an image. After several trials, we set the resolution to 1200 dpi. In the process of scanning, we also manually checked the TIFF files and applied simple image processing operations to some initial pictures like rotating pictures by 90 or 180 degrees, cropping the blank edges, etc. The benchmark is that the operations applied to the pictures do not change the attributes of the original TIFF files.

2.2 File Management

File management refers to the naming, organization, storage, and maintenance of images and related metadata. Before starting the work on scanning, we designed a naming system to name the master files. A file's name consists of the place name and a sequence number, which is more meaningful for the end users than numbers or codes. A structured directory was created to store the scanned pictures. Keeping the naming and storage consistent is very important, especially when multiple people work on the image creation. Also, it is important to have more than one copy of the scanned pictures stored at different physical places. You never can predict what may happen to the computers.

2.3 Image Delivery

Image delivery is a process of creating derivative images and getting them to the end users. To create a system that effectively supports users, it is essential to examine

users' needs and preference and to understand how the digital library is going to be used [8]. User studies have concluded that researchers expect fast retrieval, acceptable quality and complete display of digital images [6]. Although the TIFF format is very detailed, it is too big for users to download or print. For fast web delivery, the reference and thumbnail images were derived from the TIFF files and stored in the format of JPEG, which is a lossy compression format [9]. The quality of the compressed files is reduced, but the size of the files is much smaller; image quality is acceptable and it is well suited for the display of a low-resolution screen.

3 Metadata Infrastructures

Metadata is a structured description of an object or collection of objects. In the 1960s, Machine Readable Cataloging format, known as MARC format, was developed at the Library of Congress. It has twofold purposes: one is to represent rich bibliographic descriptions and relationships between and among data of library objects; the other is to facilitate sharing of the data across local library boundaries [10]. With the exponential growth of the Internet in 1990s, the MARC standard is not applicable to represent the large amount of digital information because it is not easy to identify the cataloging unit in the heterogeneous environment of the Internet as in the physical world. Also, application of the complex and detailed MARC format is very expensive. In order to bring order to the cyberspace, different communities developed different metadata schemes, describing and sharing disciplinary specific information.

As structured information repositories on the Internet, digital libraries provide an open environment for sharing information especially visual images. Digital libraries' ability to search and browse the digitized items through various access points could greatly enhance the use of the images [11]. Supporting these activities is a host of metadata schemas. Image metadata research in the Web environment is still in its infancy, and has primarily focused on the development of image-specific metadata schemes [12]. In practice, Dublin Core, VRA Core, REACH, CDWA, EAD, and IMS Learning Resource Meta-data schemes, etc. all can be used for visual images in different contextual domains.

3.1 Dublin Core as the Starting Point

In the prototype, Dublin Core metadata schema was selected as the starting point for cataloging the image items because it has been used most widely in recent years to describe electronic resources [13]. Dublin Core is regarded as a compromise between highly structured metadata and simpler models for resource discovery on the Internet [10]. It comprises a simple generic set of 15 elements applicable to a variety of digital object types. Each element is optional and repeatable. Qualifiers can be used to refine the meaning of an element or identify schemes that aid in the interpretation of an element value [14]. Furthermore, it allows for enhanced semantic metadata to be embedded into online resources such as HTML and XML, and would also provide a

format that might be able to be used to map between different, more complex metadata formats [15].

Two local considerations also contributed to the selection of the Dublin Core format—compatibility, and simplicity. In regard to compatibility, the Digital Library Center of the University of Tennessee wants to ensure the interoperability with the related Albert 'Dutch' Roth Digital Photograph Collection and Western North Carolina Heritage Collection because they are all members of the digital collection of the Great Smoky Mountains Regional Project, a cross-regional collaborated project committed to organizing all kinds of materials of the Great Smoky Mountains regions of East Tennessee and Western North Carolina. Since Dublin Core was adopted as the record structure for both of them, applying it to the Derris collection would fruitfully support interoperation and data exchange between these collections. In regard to simplicity, it requires that collections can be easily accessed by a variety of users, without requiring special browsers or plug-ins and can be easily cataloged by collection developer, without much catalog training.

3.2 Enhanced Dublin Core Metadata Schema

Tony Gill has used the term “cultural infodiversity” to express the necessarily heterogeneous nature of cultural information, with the conclusion that no single metadata schema could fit all requirements of different collections [16]. In real life projects, information professionals often need to adjust standards to fit the specific characteristics and requirements of the specific collection. In the Derris prototype project, the metadata element design is primarily restricted by the visual image nature of the original slides collection. Unlike text sources whose inherited metadata could be mined from the resource itself, there is almost nothing in the resource itself that can be mined with visual images. Most research on mining image data so far deals with pattern recognition and provides nothing in the way of content analysis [17].

An analysis of the cultural and historical image collections in other academic libraries yielded a preliminary metadata set with nine basic Dublin Core elements. In the process of cataloging, we found that the basic Dublin Core elements are not sufficient for representing the Derris digitized image collection. Some of the Dublin Core elements are not applicable to the Derris collection, like “Language”; meanwhile some additional elements are needed to describe the special attributes of the collection. Based on the further research of the collection, we developed an enhanced Dublin Core metadata schema specifically for the Derris Digital Image Library. Instead of using the names of DC elements, local field names were assigned for better understanding of the meaning they represent in the context. Totally, we defined 18 metadata fields, among which 12 elements have the corresponding DC elements, 6 are unique to the Derris collection. A record example is provided in the Appendix.

Another problem arising in the practice is that the original slides of the Derris collection provided limited descriptive information. Furthermore, the information provided by the photographer for the individual slide is not equal. Some slides have title, region, date, and the names of people in pictures, but some have no descriptive information at all. In the cataloging practice, we tried to keep as much original

descriptive information as possible. Meanwhile for the image discovery, we defined five required metadata fields so that each image could be discovered by search engines or by browsing. The five required fields are *Title*, *Region*, *File name*, *Subject* and *Call number*.

The challenges in building a digital library are more than just technical. To create a system that effectively supports users, it is essential to examine the user's needs, preferences, and work context [18]. The identification of the metadata elements, as an essential part of building digital libraries, is also directed and restricted by users needs. Since it is developed and hosted by an academic library within the university, the Derris Digital Image Library primarily serves the research needs and interests of the academic community. At the foundation of such a system must be a robust metadata schema capable of both supporting the needs of all the disciplines that comprise the university and interoperating with learning management systems throughout the university. The user study conducted by the Penn State University library indicates that the potential users of a digital image library, particularly faculty members, are more concerned about content-related issues than they are about retrieval-related issues; they are less concerned with how to discover images than with whether the image library will contain relevant images at all [17]. In the prototype, for better usability, two special fields, *Season* and *Orientation*, were designed to facilitate image discovery.

Meanwhile, because the collection will be loaded on the Internet without access restriction, it is open to the general public who are interested in learning the history, culture, mountain views, literature, art, life, economic development, sociology, human geography and anthropology about the Great Smoky Mountains region of East Tennessee. The library administrators and the digital image collection developers are also the potential users of the collection. Different metadata types support different functions and serve different purposes. In order to satisfy the diverse needs of user groups, metadata element identification must take into account the users' purposes of using the collection in addition to how they use it. In terms of the functions and the purposes they serve, the metadata elements in the prototype could be mainly divided into three classes: descriptive metadata, administrative metadata, and technical metadata. A single element that supports more than one function falls into more than one class.

- Descriptive metadata is for discovery and identification of information resources. At the local level, it enables searching and retrieving; at the Web level, it enables users to discover the collection as a whole [6]. The descriptive element in the prototype includes *Title*, *Collection*, *Photographer*, *Date of Photograph*, *Season*, *Region*, *Notes*, *File name*, *Subject* and *Call number*.
- Administrative metadata facilitates both short-term and long-term management of the digital collection. The fields *Image No*., *Format*, *Type*, *Rights* and *Call number* belong to this class.
- Technical metadata is about the image creation process and technical characteristics of the digital images. It provides information on the source data, scanner type and model, resolution, compression and file format. The technical elements are *Transmission data* and *Format*.

3.3 Authority Control of Metadata Values

In the prototype context, metadata value refers to the content or terms entered into a metadata element. For example, if the element is Date, the string "1999-6-15" would be a value. Usually, assigning value to metadata elements is subject to the ambiguity of language, the variety of conventions, and the subjectivity of personal preferences. In the visual image domain, the representation of images is further confined by the subjectivity of personal perspectives and focuses. Authority control and content guidelines are needed to ensure consistency of access points and subject headings, given the variability of the values entered into elements. A good metadata schema must provide for a method of qualifying the information held within a particular element [19]. In the prototype, two kinds of authority control tools have been applied: one is a set of standards that control the format of the values entered into the metadata elements, and the other is thesauri that control the values of the subject headings and region names.

Data Format. Three internationally agreed-upon standards were used to control the values of elements *Date of Photograph*, *Format*, and *Type*, which correspond to the DC elements Date, Format and Type respectively. In consideration of compatibility and interoperability, standards recommended by Dublin Core Metadata Initiative (DCMI) are used to control the format of the values of the three elements. According to DCMI, the recommended practice for encoding the date value is defined in a profile of ISO8601 and includes (among others) dates of the form YYYY-MM-DD [20]. The *Date of Photograph*, defined as: the creation dates of the original pictures from which the slides and digital images were derived, was decided to take the format of YYYY-MM-DD. The other consideration to adopt this format is for international usability. If the date, for example, is 5/6/79, the American people would understand it as May 6th, 1979 while the European people would regard it as June 5th, 1979. The YYYY-MM-DD format is less confusing.

As the official definition from DCMI, the *Format* element in the prototype means the physical or digital manifestation of the resource [21]. It can be used to help identify the software and hardware needed to display and operate the digital images. Since all the images used for manifestation are in JPEG format, "jpeg" was used as the default value for this field. Using "jpeg", as opposed to its popular abbreviated format "jpg" is for the reason that "jpeg" is the standard form specified in the list of Internet Media Types [MIME], a controlled vocabulary that defines computer media formats [20]. The meaning of element *Type* is the same as it is in the DC. It describes the genre or nature of the content of the resource. Since the resource of the prototype is a collection of slides derived from photographs for long-term preservation, it is a symbolic visual representation in nature. "Image", the term defined in the DCMI type vocabulary, was set as the default value of the *Type* field. Using the terms from controlled vocabulary will help to ensure the interoperability with other similar image collections.

Other than using agreed upon standards to ensure the compatibility with other collections on the collection level, it is necessary to keep consistent the format of

individual values entered into a field on the item level. It is very helpful for collection administration. Self-defined rules were defined in the prototype to generate the values of elements *Image No.* (Identifier in the DC set) and *Call number* to keep the consistency of the values assigned to them. *Image No.* is the record number that uniquely identifies a digital image object. The value consists of the term DERRIS plus a 4-digit accession number, for example, "DERRIS0001". DERRIS is used to indicate that the image is a part of the Derris collection rather than other similar collections in the Great Smoky Mountain Project. The number of slides in the Derris collection is around four thousand. Consequently, four digits provide for future identifications as they are added. Considering that some users may want to see the original slides instead of the transformed pictures, the call number field, which has the same function as the call number in physical libraries, was set up to track the relationship between the physical slides and their digital manifestations. The value of call number is the combination of the box number and the tray number where the physical slide is located. For example B01tr01 means the slide is in tray 1 box 1.

Vocabulary control. A good metadata schema must provide for a method of qualifying the information held within a particular element. It may be useful to know that the text found within a subject field comes from a controlled vocabulary [19]. A controlled vocabulary operates by choosing a preferred way of expressing a concept and then making certain that synonymous ways of expressing the concept will be connected to the preferred terminology [22]. The ambiguity of language, existence of synonyms, and subjectivity of personal interpretations may cause different representations of the subject of the same item. Using controlled vocabulary can help users collocate all the information on a particular subject in the system.

In the prototype, the cataloging unit is a digital image so that subject headings are the expression of the subject content of the individual images. We primarily decided to use the Library of Congress Subject Headings (LCSH), but in practice we found that it is too general to describe the "content" of the images. Because most images in the Derris collection are about scenery and cultural, social, and economic life in the Great Smoky Mountains, the scope of the subjects conveyed by the pictures is very narrow and very specific; Using LCSH would not help users with the source collocation in this case. Vocabularies that are developed in-house may be more meaningful and relevant for representation and retrieval. Because its sibling collection, Roth "Dutch" Photograph Digital Collection is quite similar to it in terms of the subject coverage [23], we turned to the subject headings list developed specifically for Roth collection. In the process of cataloging, according to the principle of "literary warrant", we modified some subject headings, added some terms unique to the Derris collection and finally developed an in-house thesaurus for Derris collection.

The other element that needs value control is *Region* (element Coverage in DC), the name of the place that the image depicts. Because a place may have different names or the name has been changed over time, it is not good for collocating all the images about one place if no control is imposed. Allen Coggin's book Place Names of The Smokies was used as the reference to keep values in the region field consistent.

4 Conclusions

This project provided a case study in building a digital image library based on a slide collection about Smoky Mountains' culture and history. The main steps of digitizing slides and designing metadata schema were explored. In the process of digitization, three types of images: master, reference and thumbnail were created. They have different formats and attributes to accommodate the collection requirements and users' needs. In this process, we felt that a predefined file management means is important for storing and maintaining digitized images orderly and consistently.

In the metadata design, the metadata elements were restricted by the amount of the information provided on the physical slides; however, the user-centered approach was taken to ensure the intuitive searching and browsing. Dublin Core metadata schema was chosen as the starting point for metadata design for its applicability and simplicity. An enhanced metadata scheme was constructed to represent the images based on further research of the physical slide collection and the study of users' behavior and preference. To ensure the interoperability and usability, standards recommended by DCMI were employed to control the consistency of the format of the metadata values and a thesaurus was developed in-house to keep consistency of subject headings assigned to each record. Compared to other widely used subject heading list and thesauri, the in-house developed thesaurus is more meaningful and relevant for representation and retrieval in this case. We expect that follow-up user studies would be conducted to examine the usability of the metadata design.

Asian countries have a long history and splendid culture, art, and architecture heritage, which is worth being protected and maintained for research, education, travel and entertainment. Such a large amount of cultural heritage exists in various formats and with various conditions. Digital libraries provide an economic and efficient way to preserve and disseminate the information so that the Asian culture and history could be more reachable to audiences who are interested in them. It is hoped that our experience presented here is valuable to researchers and information professionals involved in similar projects of digitization, metadata development, and integration efforts.

References

1. Garza-Salazar, David A.; Lavariega, Juan C., and Sordia-Salinas, Martha. Information Retrieval and Administration of Distributed Documents in Internet. Abramowicz, Witold, editor. Knowledge-based Information Retrieval and Filtering from the Web. Kluwer Academic Publishers; 2003; pp. 53-73.
2. Borgman, Christine L. Challenges in Building Digital Libraries for the 21st Centery. Digital Libraries: People, Knowledge, and Technology: 5th International Conference on Asian Digital Libraries, ICADL 2002, Singapore, December 11-14, 2002 : proceedings ; Singapore. Berlin, New York: Springer; 2002.

3. Cleveland, Gary. Digital Libraries: Definitions, Issues and Challenges. Accessed 2004 Jun 23. Available at: http://www.ifla.org/VI/5/op/udtop8/udtop8.htm.
4. University of Washington Libraries Digital Collections. Accessed 2004 Feb 20. Available at: http://content.lib.washington.edu/
5. Whitten, Ian H. Bainbridge David. How to Build a Digital Library. San Francisco: Morgan Kaufmann Publishers; 2003; ISBN: 1558607900.
6. Moving Theory into Practice: Digital Image Tutorial. Accessed 2004 Apr. Available at: http://www.library.cornell.edu/preservation/tutorial/contents.html.
7. The TIFF Image File Format. Accessed 2004 Feb 11. Available at: http://www.ee.cooper.edu/courses/course_pages/past_courses//EE458/TIFF/.
8. White, Martin. Information Architecture and Usability. Econtent . 2002; 25(4):46, 2p.
9. California Digital Library Digital Image Format Standards. 2001 Jul 9; Accessed 2004 Feb 19. Available at: http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf.
10. Ercegovac, Zorana. Introduction. Journal of the American Society for Information Science. 1999; 50(13):1165-1168.
11. Zeng, Marcia Lei. Metadata Elements for Object Description and Representation: A Case Report from a Digitized Historical Fashion Collection Peoject. Journal of the American Society for Information Science. 1999; 50(13):1193-1208.
12. Greenberg, Jane. A Quantitative Categorical Analysis of Metadata Elements in Image-Applicable Metadata Schemas. Journal of the American Society for Information Science and Technology. 2001; 52(11):917-924.
13. Technical Advisory Service for Images: Metadata and Digital Images. Accessed 2004 Apr 5. Available at: http://www.tasi.ac.uk/advice/delivering/metadata.html.
14. Dublin Core Metadata Initiative: Dublin Core Qualifiers. Accessed 2004 Feb. 20. Available at: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/
15. Dempsey, L., Heery, R. Metadata: A Current View of Practice and Issues. Journal of Documentation. 1998; 54: 145-172
16. Cultural infodiversity. Accessed 2004 Jun 20. Available at: http://www.rlg.org/events/metadata2002/gill/tsld003.htm.
17. Attig, John; Copeland, Ann, and Pelikan Michael. Context and Meaning: The challenges of Metadata for a Digital Image Library within the University. College & Research Libraries. 2004; 65(3).
18. Payette, Sandra D. and Rieger, Oya Y. Supporting scholarly inquiry: Incorporating users in the design of the digital library. The Journal of Academic Librarianship. 1998 Mar; 24(2):121-129.
19. Tennant, Roy . Metadata As If Libraries Depended on It. Library Journal . 2002 Apr 15; 127(7):32, 2p.
20. Dublin Core Metadata Element Set, Version 1.1: Reference Description. Accessed 2004 Jun 19. Available at: http://dublincore.org/documents/dces.
21. Dublin Core Metadata Initiative: DCMI Type Vocabulary. Accessed 2004 Jun 19. Available at: http://dublincore.org/documents/dcmi-type-vocabulary/.
22. Taylor, Arlene G. The Organization of Information. Englewood, Colorado: Libraries Unlimited, Inc.; 1999; ISBN: 1563084988.
23. Albert "Dutch" Roth Digital Photograph Collection Subject Heading List. Accessed 2004 Apr 8. Available at: http://diglib.lib.utk.edu/r/rth/sh.html.

Appendix: Example Record

Http://web.utk.edu/~yzhang/derris/record.html