Winter 2002

A look at the digitization process

Nadia Nasr
Santa Clara University, nnasr@scu.edu

A Look at the Digitization Process

By Nadia Nasr
Digitization Specialist
& SLIS Graduate Student
The University of Southern Mississippi
Nadia.Nasr@usm.edu

When I first started taking classes in library and information science, I quickly discovered that a librarian's job involves much more than ordering, processing and shelving books. I also learned more about the role of librarian as mediator between the library user and information.

Libraries traffic in information, and librarians act as gateways to that information. One of their key roles involves enabling users to access information. This was at one time limited to the written word; however, with the integration of computers into the library, this has expanded to include the virtual word as well. The World Wide Web has presented new and unexplored vistas of information, and some have said it will replace the need for librarians entirely. Instead of making librarians obsolete, the virtual world has merely presented a new challenge to librarians. The same needs for a physical library are present in the digital world: accessing materials for research, finding the locations of these materials, and verifying the authority and accuracy of these materials, as well as providing online access to research quality materials within their institution.

For the past nine months, I have had the opportunity to work with the digitization team in the Special Collections department of USM Libraries at The University of Southern Mississippi. With little more than some experience with word processing and surfing the Web, I found my work in the digital lab far more technical than I thought myself capable of. This past year, I have been involved in a team process of providing online access to research quality materials, digitizing items that are only known by the local user community and making them available to a worldwide audience. The digitization team is providing access to photographs, correspondence, oral histories and other unique items, as well as useful bibliographic information about these materials.

A great deal of work goes into the process of digitization, and in these couple of thousand words, I would like to give you a thumbnail sketch of the process of digitization, from selection of materials from the collection to making materials available for public access. I would also like to share some information about descriptive metadata.

DEFINITION

Digitization is the process of electronically scanning and converting a physical object into digital image. An image of a photograph, a letter, an oral history transcript or any other unique physical item is converted into a digital format that a computer can read. Once digitized, these images can be manipulated, sized, and enriched by a computer with software specific to this purpose and made ready for printing or mounting online.

SELECTION OF MATERIALS

The principles of selecting materials for inclusion in an online exhibit are similar to the general collection development policy of a library, with the focus on such characteristics as the intellectual value of the collection, the level of demand for the materials in that collection, and how those materials in the collection relate to other digitization efforts. Anything, literally, can be digitized, but it takes time and money and effort to do so; therefore, it is important to make certain that the materials that are digitized have value to the intended audience.

DIGITAL SOFTWARE AND EQUIPMENT

Once materials have undergone the selection phase, it is time to get down to the nitty-gritty of the digitization process, which is the actual imaging of materials. This involves an eye for taking good photographs and a little creativity. At this point computer technology comes into play.

Software and Equipment:
Software considerations include the following:

- Adobe Photoshop, or a similar image processing software
- A word processing program such as Microsoft© Word or Corel© Word Perfect
- A program that has optical character recognition capabilities for converting images of typewritten text into editable text in a word processing document
- Another optional but helpful program is software that allows the camera to be run by a computer and allows images taken by the camera to be transferred directly to the computer

Hardware considerations include the following:

- A high speed computer with a very large memory capacity, depending on the size of your project
- A CD-ROM or DVD-RAM burner for storing master images (CDs store about 700MB and DVD's store about 9.4 GB of data and both have a long shelf life)
- A high resolution monitor, at least 17 inches
- A scanner with a resolution of at least 600x600dpi
- A digital camera

With this equipment in place, digital imaging can begin. Scanners are available to image photographs; in addition, some scanners have the capability to scan negatives or slides and can be processed like any other image. For items in the collection with more than ten pages of typed text, such as an original manuscript or oral history transcript, it is recommended that these items be scanned and converted into word processed documents by the use of optical character recognition (OCR). Longer, handwritten documents can be scanned as images or transcribed by hand, but be aware that images require more storage space than word-processed documents.

Digital cameras are useful for photographing oversized items that do not fit on the scanner bed, as well as three-

dimensional objects or books whose spines are too fragile to be placed open-faced on a scanner. The selection and imaging phases of the digitization process can be really exciting, as they provide the opportunity to sift through and discover treasures that might normally go unnoticed.

DESCRIPTIVE METADATA

The next phase in the digitization process is similar to the cataloging process. Materials in an online collection need to be classified and organized in such a way that users can access the information that will best meet their needs. In the library, this is possible via the online catalog, which is a database of carefully compiled and indexed MARC bibliographic records. Describing and organizing items in a digital environment, however, requires more flexibility than the MARC formats allow. A variety of standards, such as TEI, EAD, and Dublin Core, have evolved in recent years and are designed to meet the needs for describing digitized materials.

- The **TEI** (Text Encoding Initiative) Header was originated for the purpose of developing guidelines for marking up electronic texts, such as novels, plays, and poetry, to support research in the humanities. Basic bibliographic information in TEI can be mapped to data contained in MARC, with some exceptions.
- **EAD** (Encoded Archival Description) was developed as a means of marking up data contained in the finding aids used in special collections and archives. Finding aids are sometimes lengthy narrative documents that provide information about a collection, biographical or background information, as well as information about how materials in a collection are organized and stored.
- The **Dublin Core (DC) Metadata Elements** are a set of elements or descriptors designed to allow authors to describe their own electronic resources. Simple and concise, DC is used to format data and to provide access in retrieval of an online resource, via its creator and other relevant information. A crosswalk is available for the translation of bibliographic information from MARC to Dublin Core, which illustrates Dublin Core's compatibility with MARC guidelines.

MARKUP

The next phase of the digitization process involves incorporating the images and/or text of the digitized materials and the descriptive metadata into a single electronic document that can be posted to the Web. This can be done using software that allows the digi-lab technician to 'markup' the images using an HTML editor. Dreamweaver© is a very user-friendly program with a capability for creating a basic template into which text and images can be incorporated. XML (Extensible Markup Language) is a cutting edge markup language utilizing a style sheet that converts an encoded document into HTML.

After multiple "Previews in Browser" and a few rounds of quality control to ensure that the metadata is correct and images are high quality, the collection is ready to be posted to the Web, and can be linked from the parent institution's home page.

ISSUES FOR CONSIDERATION

At this point, you have a basic understanding of the skills needed and will want to address some issues of consideration before getting started on a digitization project.

- *Copyrights:* Before deciding to publish materials from a collection to the Web, it is important to be familiar with current copyright issues related to ownership rights. It is one thing for items donated to a collection to be available within the library itself, and quite another to be published to the Web. This is especially the case in relation to personal correspondence or diaries, photographs, newspaper articles and books.
- *Condition of materials:* Are the items in the collection too fragile to be exposed to extensive handling involved in the digitization process? Digital scanning involves strong light and a significant amount of heat that can contribute to the deterioration of the item.
- *Staffing and funds:* The depth and quality of an online collection is directly proportional to the number of staff and amount of time and funds available to commit to the digitization of a collection. A small exhibit that provides a selection of items from the collection is a project that could feasibly be completed by an individual, whereas an in-depth treatment of a collection to be organized into a database will require a team of members to select and take images of the materials, create descriptive metadata, mark up the material and organize it, and then to post the exhibit to the Web.
- *Equipment and Software:* The biggest concern is the expense in purchasing and maintaining equipment and software to support the work. Research the availability of newer, high-quality versions of equipment

and software and prepare to acquire the equipment and software necessary for the project.

- *Outsourcing:* If the funds are available to do so, imaging of materials can be outsourced to an imaging provider, but keep in mind that materials will then be handled by individuals who may not necessarily know how to handle rare or fragile materials, and the quality of outsourced work is not always consistent.

AUXILIARY USES OF DIGITIZATION

Online exhibits are valuable sources of information, but if that proves to be too big a first step into conquering the digital frontier, there are some other uses of digitization that a library may find useful. Digitization is useful for providing forms online for Interlibrary Loan requests or document delivery services. With the use of Microsoft© Access or a similar software program, a database can be developed for organizing and documenting collections. There are a variety of uses for digitization in the library environment. Automation of the library catalog was just the beginning of electronically entering the virtual world.

USM LIBRARIES ONLINE COLLECTIONS:

- **Civil Rights in Mississippi Digital Archive:** This exhibit contains materials from the civil rights movement in Mississippi. This collection includes the personal diaries of such civil rights workers as Zoya Zeman and Sandra Adickes, photographs by Herbert Randall, and pamphlets and brochures from civil rights organizations, as well as oral histories of other notable figures. This collection provides access to digital images of the items, as well as their descriptive metadata. http://avatar.lib.usm.edu/~spcol/crda/

- **deGrummond Children's Literature Collection:** The de Grummond Children's Literature Collection is a research collection that contains original manuscripts and illustrations of more than 1200 authors and illustrators, as well as 70,000+ published books dating from 1530 to the present. There are four main online exhibits from this collection that showcase a selection of items from the Neubert Valentine Collection, the Kate Greenaway Almanacks, the Ezra Jack Keats Collection, and information about the Curious George Collection. Each of these exhibits showcases and describes a sample of materials from their collection. http://www.lib.usm.edu/%7Edegrum/html/showcase/sc-virtualexhibits.shtml

- **The Mississippi Oral History Project:** The Center for Oral History houses a vast collection of oral histories, and more than 800 of these are published and on deposit at McCain Library and Archives. As a result of the Mississippi Oral History Project, approximately 83 of the transcripts for these oral histories are available online and about 80 more are due to be posted online in the next few months. In the near future, sound excerpts from the oral history interviews may also be made available. Topics discussed relate to important events in Mississippi history. http://www.usm.edu/msoralhistory/

ONLINE EXHIBITS REFERENCED

- Ross, D. (2000). *Civil Rights in Mississippi Digital Archive.* Retrieved November 1, 2002, from the University of Southern Mississippi's McCain Library and Archives Web site: http://avatar.lib.usm.edu/~spcol/crda/
- Colorado Digitization Project. (2000). *Heritage Gateway to Colorado's Digitization projects.* Retrieved November 1, 2002, from the following Web site: http://coloradodigital.coalliance.org/cdp.html
- The University of Southern Mississippi. (2002). *deGrummond Children's Literature Collection* Retrieved November 1, 2002, from the University of Southern Mississippi's McCain Library and Archives Web site: http://www.lib.usm.edu/%7Edegrum/html/showcase/sc-virtualexhibits.shtml
- The Center for Oral History and Cultural Heritage and the Mississippi Humanities Council. (no date) *The Mississippi Oral History Project.* Retrieved November 1, 2002, from the following Web site: http://www.usm.edu/msoralhistory/

Online Resources

Dublin Core Metadata Initiative. (1995). *Dublin Core Metadata Initiative.* Retrieved November 1, 2002, from the following Web site: http://dublincore.org

Berkeley Digital Library (2002). *Digital Library Information Resources.* Retrieved November 1, 2002, from the following Web site: http://sunsite.berkeley.edu/Info/

Knoll, A. (1998 February). *Digitization of Rare Documents: A Solution for Preservation of and Access to Data.* Retrieved November 1, 2002, from the European Library Automation Group Web site: http://www.kb.nl/coop/elag/elag98/presentations/czech1.html

Library of Congress. (2001). *Encoded Archival Description (EAD).* Retrieved November 1, 2002, from the Encoded Archival Description official Web site, maintained by the Library of Congress: http://www.loc.gov/ead/

Colorado Digitization Project. (2000). *General Guidelines for Descriptive Metadata Creation & Entry.* Retrieved November 1, 2002, from the following Web site, maintained by the Colorado Digitization Project: http://coloradodigital.coalliance.org/glines.html

Library of Congress (2001 February). *MARC to Dublin Core Crosswalk* Retrieved November 1, 2002, from the Library of Congress Web site: http://lcweb.loc.gov/marc/marc2dc.html

Hodge, Gail. (2001). *Metadata Made Simpler (PDF format).* Retrieved November 1, 2002, from the National Information Standards Organization Web site: http://www.niso.org/news/Metadata_simpler.pdf

Cornell University Library/Research Department. (2000-2002). *Moving Theory into Practice: Digital Imaging Tutorial.* Retrieved November 1, 2002, from the following Web site: http://www.library.cornell.edu/preservation/tutorial/

TEI Consortium. (2001). *Text Encoding Initiative (TEI)* Header Retrieved November 1, 2002, from the TEI Web site: http://www.tei-c.org/