# Archiving Digital Oral History

When oral history archives consisted mostly of analog audio tapes and printed transcripts, the archivists' functions were usually seen as the final steps in the processing of interviews. Digital technology has eliminated the distinction between the creation of oral history and the preservation and management of it. Information systems must now be at the heart of the oral history enterprise, and attention to data management must begin at the moment the digital recorder is configured, even before actual recording begins. Without careful design and management of data digital oral histories cannot survive in any useful way or for any length of time.

In addition to the information provided below, read more about specific aspects of digital archive practice in Tab 4: Digitization, in this notebook.

## Archive Practice for Oral History Materials: Pre-Digital Overviews

The following books serve as an introduction to the traditional functions of archiving as applied to oral history materials. These sources introduce the major themes which concern archivists who create, preserve, document, and provide access to non-digital oral history materials: analog audio and video tapes, printed transcripts, supporting documents such as legal agreement forms and background research materials, and ancillary primary and secondary materials such as photographs, newspaper clippings, scrapbooks, and other memorabilia.

Fogerty, James E. "Oral History and Archives: Documenting Context." In *Handbook of Oral History*, edited by Thomas L. Charlton, Lois E. Myers, and Rebecca Sharpless. Lanham, MD: AltaMira Press, 2006.

MacKay, Nancy. *Curating Oral Histories: From Interview to Archive*. Walnut Creek, CA: Left Coast Press, 2007.

Matters, Marion. *Oral History Cataloging Manual*. Chicago: Society of American Archivists, 1995.

## Storing and Preserving Digital Oral History

Storing audio files, transcripts, and other digital oral history materials safely and securely for long-term preservation and access is a challenge for all oral history projects. Audio and video files are large, and the demands for digital file space grow very fast. Obsolescence of computer technology, routine maintenance and back-up, and hardware failure management all must be anticipated and planned for if recorded and collected materials are to survive for future generations.

**University of Washington Music Library and Listening Center—**
**http://www.lib.washington.edu/Music/preservation.html**
Gibb, John. "Audio Preservation and Restoration, including some links to film and video tape preservation." This Web page includes many links to sources for information on all aspects of audio and video preservation.

**LOCKSS – http://www.lockss.org/lockss/About_Us**

Copyright © 2009 by Baylor University Institute for Oral History. Permission is granted to download, print, or duplicate this document for educational purposes only.
For any other use, contact: Baylor University Institute for Oral History, One Bear Place #97271, Waco TX 76798-7271. Phone (254) 710-3437.

"LOCKSS (Lots of Copies Keep Stuff Safe) is open source software that provides librarians with an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorized content. Running on standard desktop hardware and requiring almost no technical administration, LOCKSS converts a personal computer into a digital preservation appliance, creating low-cost, persistent, accessible copies of web based content as it is published. Accuracy and completeness of LOCKSS appliances is assured through a robust and secure, peer-to-peer polling and reputation system.

"How it works:
A library uses LOCKSS software to turn a low-cost PC into a digital preservation appliance called a LOCKSS Box that performs the following four functions:

- It collects content from the target web sites using a web crawler similar to those used by search engines.
- It continually compares the content it has collected with the same content collected by other LOCKSS Boxes, and repairs any differences.
- It acts as a web proxy or cache, providing browsers in the library's community with access to the publisher's content or the preserved content as appropriate. It can also serve content by Metadata (Open URLs) via resolvers.
- It provides a web-based administrative interface that allows the library staff to target new journals for preservation, monitor the state of the journals being preserved, and control access to the preserved journals."

## Documenting Born-Digital and Digitized Oral History

Organizing, preserving, and making accessible digital oral history involves the design, creation or collection, and maintenance of metadata. Oral history materials — recordings, transcripts, and other materials — both born-digital and digitized, exist as *data* stored on some form of computer-accessible data storage device. Another level of data, called *metadata*, are the sets of information which describe, catalog, or document the provenance of the digital oral history materials and (in the case of digitized materials) the analog materials from which they are derived. Guides, standards, discussions, and training concerning archival practice for digital oral history materials all deal primarily with metadata.
An excellent introduction to metadata concepts is available online at http://www.slideshare.net/GeoffFroh/oha-2008-making-sense-of-metadata-a-practical-overview-for-oral-historians-presentation. This Web page provides access to the presentation by Geoff Froh at the 2008 meeting of the Oral History Association.

Metadata serve the following functions for oral history materials:

Discovery — information that can be searched and browsed so researchers can locate and retrieve materials relevant to their interests. Examples include traditional MARC cataloging as well as more recent and purpose-created systems for content search and retrieval.

Presentation and Navigation — information about how these materials may be accessed, with what programs and in what context, as well as aids to navigation within the digital object.

Structure — Information about how the materials are structured, in terms of both form and content, and how various disparate digital items relate to each other.

Description — Technical and narrative documentation of the provenance of the digital materials, including the source of the materials, the people, software, and equipment used to create them, and in the case of digitized materials, information about the original analog materials from which the digital objects were created. Technical documentation also serves to ensure the continued stability, integrity, accessibility, and usability of digital files, even through successive changes in technology.

Control of access and rights to a resource — information about copyright ownership and access restrictions imposed by any of the people or organizations involved in the creation or preservation of the materials.

Some metadata is embedded within digital files, either automatically by the software which creates them or by entering information into fields provided by the creation software. For example, the Adobe Acrobat program, which is often used to create archival text documents, allows for the entry of extensive descriptive and administrative metadata in the process of completing conversion of word-processing files into PDF-A1 format. Similarly, audio editing programs used to digitize analog audio recordings also allows for the entry of metadata in addition to the technical metadata generated automatically as the digital recording process is completed. There are programs which can harvest this embedded metadata for use in catalogs, finding aids, or other metadata management systems. JHOVE, a collaborative project of JSTOR and the Harvard University Library, is one such program.

**JHOVE — http://hul.harvard.edu/jhove/**

"JHOVE provides functions to perform format-specific identification, validation, and characterization of digital objects.

- Format *identification* is the process of determining the format to which a digital object conforms; in other words, it answers the question: 'I have a digital object; what format is it?'
- Format *validation* is the process of determining the level of compliance of a digital object to the specification for its purported format, e.g.: 'I have an object purportedly of format *F*; is it?'
- Format *characterization* is the process of determining the format-specific significant properties of an object of a given format, e.g.: 'I have an object of format *F*; what are its salient properties?'

"The set of characteristics reported by JHOVE about a digital object is known as the object's *representation information*, a concept introduced by the Open Archival Information System (OAIS) reference model [ISO/IEC 14721]. The standard representation information reported by JHOVE includes: file pathname or URI, last modification date, byte size, format, format version, MIME type, format profiles, and optionally, CRC32, MD5, and SHA-1 checksums [CRC32, MD5, SHA-1]. Additional media type-specific representation information is consistent with the NISO Z39.87 Data Dictionary for digital still images and the draft AES metadata standard for digital audio."

**Open Archival Information System (OAIS) Reference Model — http://public.ccsds.org/publications/archive/650x0b1.pdf**

**http://nost.gsfc.nasa.gov/isoas/**

This set of standards and recommendations was developed under the auspices of NASA Consultative Committee for Space Data Systems (CCSDS) and has been adopted by the International Standards Organization (ISO).

"This document is a technical Recommendation for use in developing a broader consensus on what is required for an archive to provide permanent, or indefinite long-term, preservation of

digital information.
"This Recommendation establishes a common framework of terms and concepts which comprise an Open Archival Information System (OAIS). It allows existing and future archives to be more meaningfully compared and contrasted. It provides a basis for further standardization within an archival context and it should promote greater vendor awareness of, and support of, archival requirements."

**Preservation Metadata**

PREMIS — **www.loc.gov/standards/premis**

PREServation Metadata Implementation Strategies has been a joint effort between OCLC (Online Computer Library Center) and RLG (Research Libraries Group, now a part of OCLC). Its objectives have been to "develop a core preservation metadata set, supported by a data dictionary, with broad applicability across the digital preservation community" and to "identify and evaluate alternative strategies for encoding, storing, and managing preservation metadata in digital preservation systems."

NISO/MIX — **www.loc.gov/mix//** and **http://www.niso.org/committees/committee_au.html**
"The Library of Congress's Network Development and MARC Standards Office, in partnership with the NISO Technical Metadata for Digital Still Images Standards Committee and other interested experts, is developing an XML schema for a set of technical data elements required to manage digital image collections. The schema provides a format for interchange and/or storage of the data specified in the Data Dictionary - Technical Metadata for Digital Still Images (ANSI/NISO Z39.87-2006)."
(http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c )

**Rights Management Metadata**

"Digital Rights Management poses one of the greatest challenges for content communities in this digital age. Traditional rights management of physical materials benefited from the materials' physicality as this provided some barrier to unauthorized exploitation of content. However, today we already see serious breaches of copyright law because of the ease with which digital files can be copied and transmitted."
Iannella, Renato. "Digital Rights Management (DRM) Architectures." *D-Lib Magazine*, 7 (6), June 2001. http://www.dlib.org/dlib/june01/iannella/06iannella.html.

## General Metadata Schemas

**Dublin Core (http://dublincore.org/)**

"The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems."
The complete list of current Dublin Core metadata terms, as well as an extensive list of links to schemas of defined terms for particular kinds of material and subject areas, can be found at http://dublincore.org/documents/dcmi-terms/#H4.

**MODS — Metadata Object Description Schema (http://www.loc.gov/standards/mods/)**

"The Library of Congress' Network Development and MARC Standards Office, with interested experts, is developing a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications. As an XML schema it is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. This schema is currently in draft status and is being referred to as the "Metadata Object Description Schema (MODS)". MODS is expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users."

## Document Encoding

Document encoding increases the *granularity* of digital resources. It is a programming tool that provides direct access to parts of a text, such as chapter or section headings, individual pages, paragraphs, or stanzas, or even individual words. Encoding can also provide links to external sources of further information, such as definitions, linguistic, historical, or other contextual information, or other resources. Several systems of text and document encoding have been developed, some for individual texts (such as a single interview transcript) and some for collections of items (such as a finding aid for a collection of interviews or a project).

**TEI — Text Encoding Initiative http://www.tei-c.org/index.xml**

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of supporting resources, including resources for learning TEI, information on projects using the TEI, TEI-related publications, and software developed for or adapted to the TEI.

The TEI Consortium is a non-profit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors."

"The TEI *Guidelines for Electronic Text Encoding and Interchange* define and document a markup language for representing the structural, renditional, and conceptual features of texts. They focus (though not exclusively) on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis. These guidelines are expressed as a modular, extensible XML schema, accompanied by detailed documentation, and are published under an open-source license." (http://www.tei-c.org/Guidelines/index.xml)

**EAD Encoded Archival Description — http://www.loc.gov/ead/eaddev.html**

"Encoded Archival Description (EAD) is an emerging standard used internationally in an increasing number of archives and manuscripts libraries to encode data describing corporate records and personal papers. The individual descriptions are variously called finding aids, guides, handlists, or catalogs. While archival description shares many objectives with bibliographic description, it differs from it in several essential ways. From its inception, EAD was based on SGML, and, with the release of EAD version 1.0 in 1998, it is also compliant with XML. EAD was, and continues to be, developed by the archival community. While development was initiated in the United States, international interest and contribution are increasing. EAD is currently administered and maintained jointly by the Society of American Archivists and the United States Library of Congress." Pitti, Daniel V. "Encoded Archival Description." *D-Lib Magazine*, 5 (11), November, 1999. http://www.dlib.org/dlib/november99/11pitti.html.

**Metadata Encoding Transmission Standard (METS) — http://www.loc.gov/standards/mets/METSOverview.v2.html**

Excerpt from "METS: An Overview & Tutorial" (URL cited above):
"Maintaining a library of digital objects of necessity requires maintaining metadata about those objects. The metadata necessary for successful management and use of digital objects is both more extensive than and different from the metadata used for managing collections of printed works and other physical materials. While a library may record descriptive metadata regarding a book in its collection, the book will not dissolve into a series of unconnected pages if the library fails to record structural metadata regarding the book's organization, nor will scholars be unable to evaluate the book's worth if the library fails to note that the book was produced using a Ryobi offset press. The same cannot be said for a digital version of the same book. Without structural metadata, the page image or text files comprising the digital work are of little use, and without technical metadata regarding the digitization process, scholars may be unsure of how accurate a reflection of the original the digital version provides. For internal management purposes, a library must have access to appropriate technical metadata in order to periodically refresh and migrate the data, ensuring the durability of valuable resources.
"The Making of America II (http://sunsite.berkeley.edu/MOA2/) project (MOA2) attempted to address these issues in part by providing an encoding format for descriptive, administrative, and structural metadata for textual and image-based works. METS, a Digital Library Federation (http://www.diglib.org/) initiative, attempts to build upon the work of MOA2 and provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users)."

## Digital Content Management Systems

Managing digital oral history requires preserving digital files in various formats, making interviews and supporting materials accessible to researchers or the general public, providing search and browse functions to make collections useful to searchers with a wide range of interests, and controlling access according to the terms of agreements with interview participants and of copyright. Ready-made and customizable systems are available to handle all of these tasks.

For an overview of some of the features required for content management systems, and some of the issues they must address, the white paper cited below may be helpful. Although it is written primarily for business enterprises, it is applicable to academic and cultural content, as well.

Rosenblatt, Bill and Gail Dykstra. "Integrating Content Management with Digital Rights Management." Giantsteps Media Technology Strategies and Dykstra Research, 2003. http://www.xrml.org/reference/CM-DRMwhitepaper.pdf

**ContentDM — http://www.contentdm.com/**

"CONTENTdm is a single software solution that handles the storage, management and delivery of your library's digital collections to the Web by providing you with:

- A Windows-based, digital collection builder where data and digital items are prepared in large batches
- A server where data and images are stored and can be edited
- A Web-based search interface and customizable display templates
- Integration with OCLC products for building collections with cataloging workflows, as well as harvesting from the Web and preservation
- Services to assist with every phase of collection development from digitization to preservation

For an example of an oral history collection deployed through ContentDM, see the Baylor University Institute for Oral History collection at http://contentdm.baylor.edu/cdm4/index_08oralhist.php?CISOROOT=/08oralhist

**Greenstone — http://www.greenstone.org/**

"Greenstone is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO. It is *open-source*, multilingual software, issued under the terms of the GNU General Public License."

**DSpace — http://www.dspace.org/**

"A groundbreaking **digital repository system**, DSpace captures, stores, indexes, preserves and redistributes an organization's research material in digital formats. Research institutions worldwide use DSpace for a variety of digital archiving needs — from institutional repositories (IRs) to learning object repositories or electronic records management, and more. DSpace is freely available as open source software you can customize and extend. An active community of developers, researchers and users worldwide contribute their expertise to the DSpace Community."