

The Problem of Ingesting and Delivering Complex Objects from Digital Repositories

Mark Kornbluh, Michigan State University
Jerry Goldman, Northwestern University
Dean Rehberger, Michigan State University

The recent emergence of online digital archives has brought educators a major step closer to bringing original, reusable digital objects into undergraduate classrooms. Yet having to search multiple archives through mind-numbing search-and-browse routines can make it extremely difficult for educators to use the repositories successfully in their curriculum. What educators need is a suite of tools that allow them to reduce the search for relevance, expand the metadata with user-specific annotation, and tie the digital libraries' content directly to course materials. The keys to creating these resources are to build distributed networks of users and repositories. Cost containment often severely limits the amount of descriptive metadata that can be catalogued. Students and instructors create topical annotated bibliographies or lists of media clips (or segments of media clips) and "publish" these for class, work group, or more general use. Allowing teachers and students to annotate and segment media as well as build their own galleries greatly enhance the educational value of digital objects by augmenting the minimal descriptive metadata and facilitating the building of complex digital objects tailored to the needs of specific education standards and curricula. The project uses a METS XML schema that provides an encoding format for administrative, descriptive, and structural metadata that is fully compliant with OAIS, and open source applications to facilitate ingestion and delivery (as well as help to control costs).

MATRIX: The Center for Humane Arts, Letters, and Social Sciences Online (<http://matrix.msu.edu>) is a pioneering research center based at Michigan State University that is devoted to the application of new technologies in humanities and social science teaching and research. All of the Center's work strives to use Internet technologies to improve education and increase the democratic flow of information throughout the world. By merging research in cutting-edge computing technologies with research in the humanities, the Center develops online educational and academic resources, provides training in computing and Internet enhanced pedagogy, and creates forums for the exchange of ideas and expertise in the humanities and in new educational technologies. Directed by Professor Mark Kornbluh, MATRIX focuses its research energies on three main challenges: the development of large-scale integrated research tools that can be developed by widely disparate repositories and freely accessed worldwide; the digitization of materials so that they can best be used by teachers, students, and researchers; and the creation of local, national and international networks that promote heritage preservation, civic education, and scholarly communities.

MATRIX is involved in a wide range of research and development projects with American and international partners including: creating online repositories and contributing to developing metadata standards for the international digital archive information exchange, digitizing materials for preservation and access, designing online training and curricular materials, and promoting meaningful international networks of scholars, activists, heritage workers, and civil servants. Currently, MATRIX is involved in a number of collaborative endeavors to digitize and make widely available archival materials, journals, artwork, artifacts, oral histories, and music for use by both academic and public audiences. The largest of these projects is the "National Gallery of the Spoken Word" (NGSW), which is funded under the National Science Foundation's Digital Libraries Initiative, Phase II program. The first large-scale repository of its kind, the NGSW is developing a significant, fully searchable, online database of spoken word collections that span the twentieth century. MATRIX also directs an International Digital Libraries project, the African Online Digital Library (<<http://www.africandl.org>>), a multilingual and

multimedia digital library currently focusing on West African resources.

Michigan State University partners with Northwestern University on many of its digital library projects. Professor Jerry Goldman directs efforts at NWU on these projects. Goldman's pioneering work in audio archives has been advantaged by the partnership at Northwestern between the Library and NUIT Academic Technologies. Goldman's work in digital audio archiving and in the use of web-enabled multimedia-based experimental research earned him coveted professional and national awards. His teaching has been distinguished for its use of novel applications including the infamous 'law-baseball quiz' <baseball.oyez.org>.

One of the latest endeavors of Kornbluh and Goldman is *The Spoken Word: New Resources to Transform Teaching and Learning**. This project is working to transform undergraduate learning and teaching through integrating the rich media resources of digital audio repositories into undergraduate courses in history, political science and cognate disciplines in the U.S. and Britain. The project takes full advantage of the flexibility inherent in digital repositories to build processes for learning that fundamentally expand the way students and teachers understand knowledge, knowledge resources, and their own complementary roles in higher education. Michigan State University, in collaboration with Northwestern University and the National Archives and Records Administration (NARA), and Glasgow Caledonian University, in collaboration with the BBC - Information & Archives, are developing and implementing this vision. Starting with a large collection of digitized audio resources, associated texts and images and a set of integrated online annotation tools, this work promotes the usability and integration of digital spoken word repositories to improve undergraduate teaching. The project tests whether and with what effect the integration of digital audio resources into university courses achieves four major project outcomes: (1) improving student learning and retention, (2) developing aural literacy in our students, (3) augmenting student competence to write on -- and for -- the Internet, and, (4) enhancing digital libraries through a focus on learning. To test these resources in the classroom, researchers at Northwestern and Matrix have developed (and continue to improve) a number of tools and processes.

The projects' archives are based on the storage archive model proposed by NASA as the Reference Model for an Open Archival Information System (OAIS). This model provides a methodology for creating an information ingestion, storage, and retrieval system that facilitates the coordination of metadata management, data management, and user/administrator feedback. It is a particularly useful model for multiple partner projects since it facilitates discussion between institutions on audience considerations, preservation techniques, and dissemination matters. It also provides a base for agreements between institutions that facilitates the sharing of information and files as well as online persistence.

In addition to work on the OAIS storage archival model, project researchers are working with the digital library initiatives at the Library of Congress, New York University, MIT, Harvard, the University of California - Berkeley, OCLC, and RLG to develop the METS (Metadata and Encoding Transmission Standard) for encoding and transmitting metadata. The METS eXtensible Markup Language (XML) schema allows for the encoding of administrative, descriptive, and structural metadata as well as extensions for documenting essential copyright information and technical data on all formats of digital media. The standard works for the goals of a digital archive for a number of different reasons. METS is an internationally known metadata standard that is being maintained by the Library of Congress and is being developed as an initiative of the Digital Library Federation. This ensures its stability and further development, both of which will facilitate long-term preservation. METS is OAI (Open Archive Initiative) compliant, allowing for the exchange of information between archives and institutions. Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP). The development for the METS extensions is based on a number of existing standards projects including NISO, MPEG7, National Library of Australia, CEDARS, NEDLIB, and LOC, among others. METS has been approved through the DLF formal review process.

Even though METS is standard enough to facilitate the exchange of metadata, it is flexible enough to incorporate both legacy metadata as well as project specific metadata.

METS also allows for different kinds of metadata to be present in one record by either wrapping or linking to other metadata. This allows it to store and common descriptive metadata standards including Encoded Archival Description (EAD), Machine-Readable Cataloging (MARC) and Dublin Core, as well as integrating isolated kinds of metadata developed by an institution. As a result of extensive research in metadata and system architecture, Matrix has made significant progress in developing searching tools that can support a variety of user-driven web-based interfaces. The Center's various content-specific archives appeal to a wide range of educators and scholars. Thus project researchers are designing multiple searching tools tailored to the specific needs and academic interests of these different types of users. Particular attention is paid to designing tailored interfaces and databases that are searched in customized ways so that information suits the needs of different users. Through the use of METS and a full-text-database, users can do general keyword searches to explore the breadth of the collections. They can also specify exact parameters to do highly refined deep searches to find the exact materials they need.

Because of the innovative design of the database and documents, users also have other ways to access the materials. Users will have an option to have their queries returned as a traditional list of relevant documents but they can also have the option to be supplied with complex objects that bring together a number of associated media (audio, images, bibliographies, teaching aids, and related documents).

The Spoken Word repository known as "repos") is a complex ingest and delivery system built with open source tools. Given the current limitations of XML databases, XML is not currently used to store metadata in the repository. However, the database structure was designed to mirror the functionality of XML with the intent to migrate to XML in the future. Repos can currently output XML for metadata harvesters. Modeled after the METS schema, the database table structure of the repository was developed to be highly flexible. Because the repository integrates so many different kinds of metadata from different institutions, the design moved away from statically defined tables and incorporated a schematized table design. This design also allows partners within the repository the flexibility to design and modify their own metadata schemes and

ingestion/administration forms for objects within the repository through an easy to use browser based utility. Because the repository stores information about the kind of metadata used by each project, MATRIX developed a PHP-based online utility that utilizes this information to generate dynamically metadata ingestion/administration forms for each of the partners. This utility allows projects to begin their participation in the repository by selecting from existing metadata schemes (Dublin Core, MARC, EAD, etc.) or entering information about project-specific metadata to describe the objects they will store in the repository. This information is then stored in the database and instantly used to generate online forms to begin entering metadata. This tools developed facilitate the dynamic generation of galleries and aid in the searching for files by format.

The repository's primary use is for storing files and their associated metadata. The supports multiple projects each designed to ingest metadata for a different type of file(s).

To accomplish our goal of supplying simple and custom forms for data entry, researcher built what has been entitled as a "Project Designer". This section of the website allows the designer to build forms that are dynamically generated for metadata entry. Each form element creates a "control," which can consist of multiple HTML form fields that work together to provide an intuitive interface for data entry. These controls are managed by a system that was previously developed called FAS which automates the process of displaying a form, validating the input, and redisplaying the form if necessary. The Project Designer offers a web interface to add controls, organize the controls into groups, and edit additional information such as control modifiers, captions, and detailed descriptions. In addition to allowing partners to build forms, they can import "Project Schemes" into their projects. If a project is marked as a scheme, then other projects can import that scheme, and all of the schemes fields will be referenced by the new project. Once the project is designed, data entry can done from any location that has internet access. Every record (file, with metadata) in repos has a unique id called a pbd (project-base-deriv) The pbd is a 3 part string consisting of the project prefix, the master id, and the derivative id separated by dashes. After metadata is entered, and submitted, FAS validates all the data. If the data doesn't validate, it redisplay the form (keeping entered values) along with error

messages. Once it validates, metadata is entered into the metadata table. After the metadata is entered, partners can upload their file. Files in any digital format can be added as well as derivative files. This process allows partners to control, edit, and maintain their collection and distribute the labor of metadata entry.

In addition to ease and distribution of data entry, Repos also allows for a variety of delivery formats including browse and search, galleries (both project and user developed), structured presentations and exhibits, and user portals that can be used for classroom presentations and assignments. Most important for educators will be contextualized search returns. For example, for developing a course in US history, an educator will use standard digital library returns of one object at a time--a speech, an image, etc. The portal interface designed for Spoken Word will return each object in context (i.e., Eisenhower speech on Domino theory with pictures of Eisenhower, other cold war speeches and images, other Eisenhower speeches, links to relevant documents, biographical data, lesson plans, and websites).

Central to the augmentation of metadata are tools that allow educators to segment and annotate audio files. One of these tools in the process of development is NoteTaker, an interactive tool that will allow user-driven segmentation and annotation of streaming audio files. NoteTaker is an easy-to-use interactive API for general web annotation and collaboration now under development at Northwestern University. It is being extended to apply to audio resources as well as to text and images. NoteTaker can be used to organize and share an outline of research notes with an instructor or others in a work group. The software provides a simple web annotation system that lets users take and organize notes linked to web based resources. The user creates an outline that is linked to web pages or objects within web pages. The user can browse the notes in the outline and view the associated web pages and resources or browse web pages and view the notes that are attached to that page. Notes and outlines are stored on a server and can be shared between individuals and groups. NoteTaker can be used to submit work to an instructor for review and comment or can allow a group of students to work collaboratively on a shared project. A new version of NoteTaker is nearing completion that allows the user to control access to notebooks,

allows threaded discussions, and supports the editing and storage of structured content (forms).

NoteTaker facilitates a process of marking and saving user-specific timestamp and metadata information. Users can listen to long files of audio, select particular segments of interest, and mark beginning and end points for those segments within the digital audio stream. NoteTaker also facilitates user annotation of the selected audio. Teachers, for example, annotate a selected audio segment with notes on potential uses for themselves, course-related comments for their students, or grade level information and instructional suggestions for other teachers. The tool then saves user preferences to a server, including time code data, annotation and other metadata information users may enter.

While all digitized audio used in this project has been ingested into the system with fundamental administrative and structural metadata, and basic descriptive metadata (usually Dublin Core), cost containment severely limits the amount of descriptive metadata that can be catalogued. Annotation by teachers and students can therefore greatly enhance the educational value of digitized audio in our system by augmenting this minimal descriptive metadata with their annotations. Students and instructors will create topical annotated bibliographies or lists of media clips (or segments of media clips) and "publish" these for class, work group, or more general use. Students are then invited to assign (provided) topic headings and rate materials relative to those topic headings. For example a student might rate a clip as very important for "exploring the controversy regarding African American voting rights in the 1960s." Students will be able to contribute short (two paragraph) reviews of particular media clips and segments. Every user is provided with a notebook used to collect links to media clips or segments. Shared group notebooks also let materials be shared among users. When a user listens to a media clip, she or he will be given the option to "collect" the clip in a private or shared notebook.

The annotation tools are web forms that have a variety of controlled vocabulary fields and open descriptive text boxes. Number of voices, time, grade level and subject categories, for instance, are selectable through check boxes and dropdown menus containing controlled vocabulary; however, open-field text boxes for content summary, user notes, teacher notes,

student notes would allow free-form user entry. These annotations and segmentations are saved in descriptive metadata records, in a MySQL database for translation to XML files. Each file is saved as equivalent to the other files to enable searching by any field within the record. Similarly, through use analysis, researchers can prioritize search results from the database by any field, or by popularity or user comments. Researchers can also weight certain metadata fields to order search results for different users.

Ease of use is important since educators (and students) often have little time to invest in learning new technologies. Similarly, university faculty members are often asked to use software packages adopted by their institutions. Thus, educators have come to rely on courseware products such as Blackboard, WebCT, Anlon, WebLearner, among others, which ease the creation and use of online resources for both teachers and students. These increasingly familiar environments serve as a "launching pad" for audio annotation tools developed by this project because they are being developed to be compliant with the Open Knowledge Initiative (OKI), increasing the likelihood of future alliances with university and commercial partners.

In addition to the work on metadata and system architecture, MATRIX has also made great strides in its research on educational uses for sound. To this end, MATRIX personnel have created a SMIL package that allows users to listen to recordings while viewing images and text. The automated SMIL creation application suite uses an XML (trans13.dtd) Transcriber document and a RealMedia audio file as its input and creates a time-synchronized SMIL presentation with an audio timeline and an accompanying scrolling text display in one RealPlayer window embedded into a custom HTML window. The process is entirely server-side and is controlled by a user-friendly web browser interface.

*The Spoken Word Project, a \$1.5 million grant from the US National Science Foundation (NSF) and the UK's Joint Information Systems Consortium (JISC). References can be found at <http://www.historicalvoices.org/spokenword>.