The Levy Project: A Progress Report

This is a progress report on the Levy Project, which is now in its second phase. The project involves digitization of the Lester S. Levy Collection of Sheet Music (Milton S. Eisenhower Library, Johns Hopkins University), which is a collection of more than 29,000 pieces of American popular sheet music spanning the years 1780 to 1960.

In Phase One, images of the music and lyrics, and color images of the covers of the Levy Collection were digitized and a database of text index records was created. This database is available at: `http://levysheetmusic.mse.jhu.edu`. Phase Two consists of converting the digitized music to computer-readable format along with full-text lyrics, generating sound renditions, and creating metadata to enhance search capabilities.

The entire project is an experiment in developing a comprehensive framework of tools to manage the workflow of large-scale digitization projects. This framework will not only support the path from physical object and/or digitized material into a digital library repository, it will also offer effective tools for incorporating metadata and perusing the content of the resulting multimedia objects. The Levy Collection, with its large size and availability in digital format, is an ideal subject for development and evaluation of this proposed framework.

Phase One of the Levy Project focused on digitally photographing the music into TIFF files, converting to JPEG images and thumbnails, and then mounting the images on the Web. Online indexing was also created at the sheet music item level. An index record for each piece of music title was created, which included when available: the unformatted transcription of title, statement of responsibility, first line of lyric, first line of chorus, dedication, performer, artist/engraver, publication information, plate number, and box and item number. We also introduced controlled vocabulary in the form of brief subject terms—for both the content of sheet music covers and content of songs—from the Library of Congress's *Thesaurus for Graphic Materials*. Currently the information is available as unformatted free text files that can be searched by simple keyword or phrase.

In Phase Two, optical music recognition (OMR) software, developed by one of the authors, is used to convert the TIFF image of scanned sheet music into computer readable-formats, which includes NIFF (Notation Interchange File Format) and MIDI files along with full-text of the lyrics. These digital objects will be deposited into the data repository along with the scanned sheet music TIFF, JPEG and thumbnail, and associated metadata.

Our OMR software offers five important advantages over similar commercial offerings. First, it can be run in batch processing mode, an essential feature for the Levy Collection given its large number of music sheets. It is important to note that most commercial software is intended for the casual user and does not scale for a large number of objects. Second, the software is written in C and therefore is portable across platforms. Third, the software can "learn" to recognize different fonts, also an issue considering the diversity

of the Levy Collection. Fourth, the software will be in the public domain. Finally, this software can also separate full-text lyrics that can be further processed using commercial optical character recognition software.

Both output of the OMR, NIFF and MIDI, represent well-recognized standard file formats. NIFF is a file format designed to allow the interchange of music notation data between and among music notation editing and publishing programs and music scanning programs. MIDI provides low-bandwidth transmission of music over the Internet so that the users can listen to the music with their web browsers.

To enable powerful search and retrieval as well as user-friendly navigational mechanism, Phase Two of the Levy Project will include a strong metadata component. Commonly defined as "data about data," metadata is structured representational information. The kinds of metadata important for Levy include descriptive (to enable searching, browsing and identification of items), structural (to enable the creation of an interface for optimum browsing and navigation), and administrative (to manage the digital components of the collection and aid users in identification of items).

The current index-text will be converted into structured metadata using XML (Extensible Markup Language) tagging and tied or "bound," perhaps as a type of header akin to the TEI (Text Encoding and Interchange) header in wide use with marked up texts, into the digital sheet music along with the image, sound, and text versions. We will use the XML markup to create indexes that will allow users to move between general keyword and precise searches. We will extract rich name information from the unstructured index-text into specific indexes such as composer, lyricist, or arranger, and possibly performer, artist, engraver, lithographer, dedicatee, and publisher. Cross-references will direct searchers to index records containing varying forms of names, including pseudonyms, transcribed from the sheet music pieces. All records will have the "authoritative" version of names. The subject terms will also receive mark-up to facilitate subject keyword searching.

To further enhance the scholarly value of the Levy Collection, a web interface will be developed for a music research toolkit (e.g., David Huron's Humdrum). These toolkits are software tools intended to assist in music research and are suitable for use in a wide variety of computer-based musical investigations, such as motivic, stylistic, and melodic analysis and concordance studies. We also propose to extend plans for developing automated means of mining authoritative name information and creating even richer name indexes. Together these added components will extend the foundation already in place for exploiting the scholarly riches of the Levy Collection. Researchers will be able to take advantage of keyword searching across index-texts, viewing the digitized sheet music and covers, linking to full-text lyrics, and hearing sound files. While adding sound files to Levy will be in itself a major achievement, the project will provide users the means to use the sound files not merely for the sake of hearing the music, but also to perform sophisticated searching. These improvements will increase significantly the collection's research value to scholars, educators, writers, musicians, and the general public.