# PRESERVATION DIGITIZATION OF DAVID EDELBERG'S HANDEL LP COLLECTION: A PILOT PROJECT

**Catherine Lai** **Beinan Li** **Ichiro Fujinaga**

Music Technology, Faculty of Music
McGill University
Montreal, Canada

lai@music.mcgill.ca beinan.li@mail.mcgill.ca ich@music.mcgill.ca

## ABSTRACT

This paper describes the digitization process for building an online collection of LPs and the procedure for creating the ground-truth data essential for developing an automated metadata and content capturing system.

**Keywords:** Digitization, Preservation, Analogue Sound Recordings, Use and Access, Digital Library Collections.

## 1 INTRODUCTION

Long-playing phonograph records (LPs) were one of the major analogue recording formats distributed commercially throughout most of the twentieth century. Although most of these historic sound recordings have long shelf lives, compelling reasons have led to a shift toward digital preservation.

To assure preventative preservation and facilitate new forms of access to this very important cultural heritage, a large digitization effort is required. An efficient and economical workflow management system is essential to carry out the steps in the digitization process. This digitization process is time-consuming and expensive since many steps involved in the digital conversion, such as metadata extraction, require much human intervention and a high-level musical and bibliographic knowledge.

It is essential to minimize human intervention so as to reduce the cost of digitizing very large numbers of LPs. One way of achieving this is to integrate sophisticated pattern recognition systems to automatically generate text and metadata from the captured images. Another time-consuming task, if performed by a dedicated human digitization operator, is separating the music tracks that are on each side of audio discs. A plausible approach to automating track separation is to use digital signal classification techniques.

Approximately thirty LPs from David Edelberg's Handel collection were digitized as a pilot study. The LPs, housed in McGill University's Marvin Duchow Music Library, are one of the largest collections of analogue recordings of Handel's music. Much of the effort at this initial stage of the project was devoted to digital benchmarking for conversion and access and to creating ground-truth data that can be used to train and test content analysis systems, thereby automating the digitization process.

## 2 BACKGROUND

Digital library projects focusing on audio preservation are still in the development stage. The Loeb Music Library Audio Preservation Studio of Harvard University is currently examining the methodologies and technologies needed to access sound recordings and other digital objects [1]. The Digital Audio-Visual Preservation Prototyping Project of the Library of Congress (LC) is investigating approaches for reformatting recorded sound and moving image collections, with a focus on metadata [2]. The University of California at Santa Barbara is conducting a pilot project on cylinder preservation and digitization [3]. Other related research projects on sound recordings include Indiana's Variations2 project [4], the digitization of 78rpm recordings at the Frontera Archives [5], and the Digital Audio project at the National Library of Canada [6].

The digital preservation of the Edelberg Handel collection is unique for several reasons. It deals with a large collection of LPs, involves digitization of both audio and visual components (album covers and liner notes), and involves benchmarking for conversion and access. It implements an integrated database with searchable full text, images of album covers and record labels, and audio files of LPs. Furthermore, it develops automated content capture systems to reduce the cost of digital conversion whenever possible.

## 3 PREPARATION OF THE QUALITY CONTROL ENVIRONMENT

Quality control (QC) is an essential and integral component in various stages of digitization. The quality of digital reproduction rests to a significant degree on the QC instruments and software [7]. The Handel digitization project uses state-of-the-art digitization equipment and software tools to reformat and reproduce analogue sound recordings. The multimedia digitization workstation consists of professional models of a record cleaning machine, turntable, and large-format flatbed scanner; a phono-preamplifier and A/D audio converter; and a powerful workstation with dual monitors for close visual inspection of image quality. Other QC features relating to equipment installation, configuration, and development include colour management, which ensures colour consistency from image capture through display, and monitor calibration/optimization, which ensures onscreen accuracy by setting white and black point, gamma, and colour balance.

# 4 COPYRIGHT AND RIGHTS MANAGEMENT

Copyright of sound recordings usually belongs to the record labels that issued the recordings, and it is the single most important legal issue to consider when planning a digitization project. However, since LPs are often the aggregate creation of several parties, individuals such as photographers, translators of lyrics, designers of artwork on album covers, and others also have varied rights to the use of the sound recordings. Since the dates of LP release were not widely indicated on LP labels or album covers before the inclusion of phonogram dates beginning in 1976 [8] and renewed or extended copyrights due to re-release of album records can occur, legal clearances of LPs for digitization require complex rights management checks via various sources such as the WorldCat OCLC Online Union Catalogue, the Bielefelder Catalogue, the Diapason Catalogue, the Schwann Catalogue, and the Gramophone Catalogue.

# 5 METADATA SCHEMA

## 5.1 Need for a Metadata Schema

Traditional standards for cataloguing sound recordings exist [9, 10, 11] but are inadequate for digitized representations of LPs as cultural heritage objects. The standards are generally limited to bibliographic description of relatively few elements [12]. Information about artwork or photographs in the album packaging, for example, is usually not included.

A metadata schema was designed to better facilitate the management and use of resources from the intellectual content to the research of rights and permissions. This schema was created to the finest level of granularity possible in order to meet user needs and provide various functionalities in the context of a digital library.

## 5.2 Existing Metadata Standards

Various efforts in the archival communities, digital library research groups, and museums have established recommended metadata element sets for multimedia such as digital still images. A few well-known metadata standards designed for visual collections are the Categories for the Description of Art (CDWA) [13], the Visual Resources Association's core categories (VRA Core 3.0) [14], Harvard's data dictionary of administrative metadata for digital still images [15], and the National Information Standards Organization's Data Dictionary of technical metadata for digital still images (ISO Z39.87) [16].

Other recognized authorities have contributed and expanded the utility of metadata element sets to digital library service models. The Making of the America II Testbed Project of the California Digital Library (CDL) worked extensively to identify and define the structural and administrative metadata elements that are crucial in the development of digital library services and tools [17]. LC also developed a set of core metadata elements, including administrative, structural, and descriptive metadata, for the LC Digital Repository Development [18].

Metadata standards for describing sound recordings, in comparison to digital still images, are relatively few. MARC21 has been used for bibliographic description of sound recordings in major libraries, but the traditional cataloguing practices take a forced one-schema-fits-all approach. Some metadata schemes (e.g., the core metadata elements of LC) have a very small number of elements pertaining to characteristics of sound recordings: audio bits per sample, audio channel configuration, audio channel information, and audio sampling frequency. MPEG-7 [19] defines elements for description of audio and video content. However, none of these existing standards have characteristics tailored to the structural complexity that is necessary for the full description of sound recordings, such as information at the individual track or song level (e.g., recording session date, performers, recording engineers, recording equipment used, etc.).

## 5.3 Metadata Schema Design for Sound Recordings

A comprehensive metadata schema for describing LPs has been designed to facilitate resource discovery, rights management and access control, as well as administration and preservation in the networked environment. The schema includes five types of metadata: description metadata to enable discovery and identification of resources; administration metadata to support management of resources; structure metadata to describe font and layout characteristics of texts; legal rights metadata to protect intellectual property rights; and technical information metadata to record the capture process and technical characteristics of digital objects. The metadata type and categorization designs described above are based on informed decisions believed to best serve the delivery, organization, and management of networked information relating to sound recordings. The new metadata schema provides for complete auditory, pictorial, and textual content analysis. Characteristics from Dublin Core, MARC21, MODS, METS, TEI, metadata schemes dedicated for visual collection and audio as introduced above, and others were partially incorporated into its design. The current schema contains more than 120 fields.

### 5.4 Examples of Metadata Elements

Metadata is created at different levels to facilitate the management of the wide variety of components (e.g., tracks, discs, performers, recording sessions, etc.) and combinations of these components that comprise sound recordings. Metadata belongs to one or more of the following hierarchical classes: Collection, Album, Image, Disc, and Track. Examples of metadata elements at different levels are:

- Collection level: summary, subject, scope and content.
- Album level: title, varying form of title, language of title, series statement, label name, form of musical composition, acknowledgement, handling instructions.
- Image level: description, date of photo, photographer.
- Disc level: audio disc size, number of tracks, playing speed, matrix numbers, duration, playback channel.
- Track level: title, duration, date of work, performers, date of recording, recording location, recording equipment used.

## 6 CONTENT MANAGEMENT

### 6.1 Web Data-entry Form

A web data-entry form written in PHP was implemented for the encoding of LP data and metadata using the metadata schema. The metadata entry system enforced quality control, using check boxes and option buttons whenever possible to reduce typing errors. The form also incorporated dynamic features, allowing multiple entries of one metadata element (e.g., tracks). For data-entry fields that stay unchanged throughout one entire digitization session (e.g., scanning equipment), the form provided auto-fill options to populate the repeating data values. Moreover, the form employed error checking to validate data before submission to a relational database. The data-entry form also provided easy-to-update features to modify existing records stored in the database.

### 6.2 Database Design and Maintenance

A relational database in MySQL was designed and implemented to hold the metadata and the content of the digitized material. The database model reflected class hierarchies presented in the metadata schema designed for sound recordings.

Data verification and database tuning were conducted iteratively throughout the digitization process for quality assurance and performance considerations. This guaranteed data accuracy and helped to attain transparency in database design.

## 7 DIGITIZATION

The digitization process began with cleaning each disc and digitizing audio at 24bit/96kHz using an audiophile-quality turntable and a cartridge. This was followed by scanning all images, including the album covers, audio discs (for labels and matrix numbers), and any accompanying materials at 24bit/1200dpi. The process finished with metadata entry and text conversion.

### 7.1 Audio Capture

An LP must be as clean as possible to achieve optimum audio quality; therefore, each side of the records was thoroughly vacuum-cleaned before each audio digitization to remove any dirt or surface particles. The cleaning and digitization of one side of an audio disc took approximately 30 minutes.

### 7.2 Image Scanning

Image digitization of LPs included scanning the album covers, audio discs, and all accompanying material. The LPs were in the standard size, i.e., 12 inches on each side and 12 inches in diameter for the covers and discs, respectively. To ensure consistency and improve the exchange of images across a wide range of display equipment, a small colour separation guide (Kodak No. Q-13) composed of a set of standard colour patches (primaries, white, and black) was always placed in the same position relative to the scanned objects. Scanning an image at 24bit/1200dpi took approximately 13 minutes and saving the file to disk took an additional 12 minutes.

### 7.3 Metadata Extraction and Text Conversion

Metadata extraction and text conversion were two of the most expensive steps of digitization. This included measuring the physical positions and size of the visual contents of the album and any accompanying material. High-level musical and bibliographic knowledge was required to extract and enter metadata using the web data-entry form, and text conversion of program notes on album covers or any accompanying material required typing columns of text that took at least 20 minutes per column of approximately 670 words.

An average of six hours was needed to process a phonograph album using manual entry in this initial experiment. Although taking the physical measurement was extremely time consuming, it is estimated that even without this requirement, the process would still take about three hours per phonograph album.

## 8 CREATION OF DIGITAL MASTERS AND DERIVATIVES

Although there are several standards governing the creation and use of digital images and audio, such as CDL's digital image format standards [20] and the IASA Guide [21], there is no uniform approach that suits all circumstances [7]. To ensure fitness in different situations for the creation of digital images and audio for various purposes (i.e., archiving, accessing, or browsing), different technical specifications were examined and developed for the digitization of the David Edelberg Handel LPs.

### 8.1 Requirements and Evaluation of Digital Images

Two representative items in distinctive form were assessed to realize a scanning template for the creation and use of digital images: an audio disc (made of vinyl) and its album cover (made of cardboard). Requirement definitions were developed to set the evaluation criteria for the scanned images, and these included attention to the legibility of the smallest text, preservation of colour appearance, and speed of delivery.

The strategy pursued in the creation of the digital masters was matched to the technology available at hand. Michael Ester has argued for the creation of rich digital masters to safeguard the long-term value of images and the investment in acquiring them [22]. The focus of the archival copy was therefore on fidelity in addition to legibility. To exploit the best resources (e.g., storage space and computing power) and utilize the current state-of-the-art scanning technology (e.g., scanning resolution and depth), master files were created at 24bit/1200dpi. For the file format of the preservation copy, two lossless compression formats, TIFF and PNG, were examined. Although TIFF has been the preferred format for long-term retention in the digital library community, some groups, such as the Technical Advisory Service for Images, favour PNG and other formats such as SPIFF because they are open formats, offer good metadata, and use better lossless compression [7]. A comparison of the file size among non-compression TIFF, lossless compression TIFF (i.e., TIFF LZW) and lossless compression PNG was made. The significance in file size reduction (e.g., 863MB vs 400MB vs 337MB) informed the decision to save digital masters in the PNG lossless compression format.

Access dictates other evaluation criteria. Three types of access files were automatically generated: a print file (300dpi), a web display file (96/120dpi), and a thumbnail file (72dpi). The requirements for the access file are to meet legibility of the smallest text when displayed online, preserve colour appearance, and meet a reasonably fast speed of delivery. A good access image conveys the most desired information given constraints such as the speed of delivery and user tolerance. Two different specifications were used for the creation of access files due to the inherent differences in the nature of the two types of material. For the presentation of album covers, it was determined that the smallest text was becoming illegible on a standard computer screen at resolutions less than 24bit/96dpi. A resolution at 96 dpi was therefore chosen for the access files of album cover images while 120dpi was chosen for the label of the audio disc which often contained very small texts.

Two lossy compression formats, JPEG and GIF, were also examined. JPEG was preferred over GIF because album covers often contain art that consists of many colours, and GIF is usually recommended for compressing graphics that have large areas of the same colour.

The specification for the thumbnail files was based on another set of evaluation criteria. Since visibly rich data is unlikely to be perceived in this browsing format, the emphasis was placed on ease of identification rather than detailed viewing. The technical specification of the thumbnail files for scanned album images was determined to be at 8bit/72 dpi. The same specification was used for the thumbnail files of the audio disc. The GIF format was chosen over the JPEG format for thumbnail files because of its generally smaller size.

Special consideration was given to the creation of digital masters and derivatives of scanned audio discs. Specifically, each audio disc, as image, was scanned and saved in its entirety during archiving. However, for better display purposes, the access and thumbnail files of audio discs included only the label of the disc, which was cropped from the original scanned images.

### 8.2 Batch-processing of Image Derivatives

The process of creating different versions of images from master files was accomplished automatically using the open-source, cross-platform software ImageMagic. Using UNIX shell scripting, the colour separation guide that was included in the original scan was cropped and different versions of derivatives were created. The time to convert a 1200dpi TIFF image of approximately 860 MB to PNG format was about 35–50 minutes, and the time for JPEG format was about 30 minutes on a 2.8GHz Pentium 4 computer running RedHat Linux 9.0.

### 8.3 Requirements and Evaluation of Digital Audio

There are very few guidelines for audio digitization requirements and the creation of audio derivatives. The audio-visual prototyping project of LC is one of the few research projects that has put forward recommended technical specifications for the master and derivative service files [2]. IASA is another research group that has published guidelines on the production and preservation of digital audio [21].

For the audio digitization of the Handel LPs, preservation masters files were captured at 24bit/96kHz in mono or stereo, depending upon the characteristic of the LP, and saved in the industry standard AIFF format. Access or derivative service files were created in various levels of fidelity in different formats: WAV files of 16bit/44.1kHz, MP3 files of 192kbps and 112kbps, and Ogg Vorbis files of quality 5. WAV files were created to offer high-quality audio that can be easily used on both Macs and PCs. Access files in MP3 format were created in higher and lower fidelities to provide good-quality audio with smaller file sizes. Ogg derivative files were created because the format is completely free, open and unpatented. This format is also currently gaining popularity and is included in popular media players under both Windows and Mac OS X. Furthermore, Ogg quality 5 is the official setting recommended for representing music of CD quality [23].

### 8.4 Batch-processing of Audio Derivatives

The process of creating different versions of audio derivatives was accomplished automatically using the open-source, cross-platform software sndfile-convert, SoX, and LAME. The first step in audio derivative creation was to convert the original 24bit/96kHz AIFF files to 16bit/96kHz AIFF files, which took approximately 50 seconds for one side of an LP disc (about 25 minutes). The second step was to convert the audio sampling rate to 44.1kHz, which took approximately 12 minutes. Derivative files of lower fidelities were subsequently created (i.e., Ogg quality 5 and MP3 of 112kbps and 192kbps). SoX was used to derive Ogg Vorbis files from WAV files and LAME was used for WAV to MP3 conversion. The time needed to convert the specified Ogg file from a 16bit/44.1kHz WAV file of approximately 25 minutes of music was about 4 minutes, and the time used to convert 112kbps and 192kbps MP3 files was about 3 and 5 minutes, respectively.

## 9 WEB DELIVERY

A web site written in PHP was designed, developed, and maintained to facilitate easy information retrieval of the digital collection of David Edelberg's Handel LPs. This site provides online access to the intellectual content and reproductions of both images and audio. All the metadata associated with an LP, links to separated and continuous audio tracks, and scanned images of album covers and any accompanying material are displayed for each record. Multimedia files in different formats and resolutions are offered to meet diverse user needs due to variations in computer platform and connection bandwidth.

## 10 CHALLENGES

The structural complexity of music and LPs imposes many challenges to developing digital collections of phonograph recordings. One of the challenges in building a digital collection is to define metadata for sound recordings and determine the level of detail to maintain the various kinds of metadata [7]. Defining the level of granularity for the metadata is important and challenging because the success of digital preservation efforts rests to a significant degree on the scope and completeness of the metadata recorded.

Because Handel LPs are available in different languages in addition to English, and they also contain pieces by other composers, another challenge in building the digital collection is to develop and adopt a name authority control. Such control would allow management of variations in spelling (e.g., names of composers, performers, producers) and musical work titles. Similarly, a systematic vocabulary control is necessary for proper semantic description of image content of artwork on album covers for effective search based on image description keywords.

Other challenges in developing a digital collection include evaluating the effectiveness of the overall workflow management system and testing the usability of the initial model of the web delivery system. The evaluation of the digital derivatives, for example, was entirely based on visual perception. A standardized evaluation guideline based on concrete empirical evidence is necessary to determine the best practices.

Lastly, the complex rights management for different elements of LPs, including photographs, artwork, trademarks, music, music arrangements, lyrics, etc. remains a complicated task.

## 11 FUTURE WORK

An immediate task for this project is to automatically separate music tracks using digital signal classification techniques. An open-source software to remove pops, clicks, and other noise from LP recordings is also under development. Another challenging future task is to automate the metadata and content extraction. This goal can be achieved by using the ground-truth data already captured in this project and further developing the structured document analysis application, Gamera [24]. Finally, software and tools to automatically link bibliographic records from existing MARC records during metadata entry will be developed, thus further reducing the cost of metadata extraction.

## 12 CONCLUSION AND SIGNIFICANCE

Due to the enormous quantity of existing recordings and the time required to properly and faithfully digitize them, an efficient and economical workflow management system for digital conversion is necessary. This project is part of a larger research plan to develop frameworks and tools for creating distributed digital music archives and libraries. By developing digital collections of analogue holdings such as David Edelberg's Handel LPs, libraries protect their special holdings as part of a cultural necessity and meet the preservation goal by limiting physical access to the original sound recordings. With full information capture using advanced digital technology, moreover, libraries increase the serviceability of their underutilized collection with improved access points. This allows new research and educational uses of valuable but rare holdings, delivers information content directly without human intervention, and facilitates simultaneous use by potentially competing users across traditional boundaries. By applying and using emerging post-processing techniques and tools, libraries can offer unprecedented services, presenting audio and visual content in optimized digital quality that is otherwise impossible to achieve with the original sources.

## ACKNOWLEDGEMENT

This research is funded in part by the "Richard M. Tomlinson Digital Library Innovation and Access Award," David Edelberg Foundation, CIRMMT, and the FQRSC. The authors would like to thank Rebecca Fiebrink and Cory McKay for proofreading.

## REFERENCES

[1] Eda Kuhn Loeb Music Library. Music from the archive. Retrieved July 23, 2005 from http://hcl.harvard.edu/loebmusic/audioproject.html

[2] Library of Congress. Digital audio-visual preservation prototyping project. Retrieved July 23, 2005 from http://www.loc.gov/rr/mopic/avprot

[3] University of California, Santa Barbara. Cylinder preservation and digitization pilot project. Retrieved July 23, 2005 from http://www.library.ucsb.edu/speccoll/pa/cylinders.html

[4] Dunn, J. "Beyond Variations: Creating a digital music library," *Proceedings of the International Conference on Music Information Retrieval,* 2000.

[5] Association of Research Libraries. Sound savings: Preserving audio collection. Retrieved July 23, 2005 from http://www.arl.org/preserv/sound_savings_proceedings/diamant.html

[6] Library and Archives of Canada. Digital audio at the National Library of Canada. Retrieved July 23, 2005 from http://www.collectionscanada.ca/9/1/p1-248-e.html

[7] Kenney, A., and Rieger, O. *Moving theory into practice: Digital imaging for libraries and archives.* Research Libraries Group, Mountain View, CA, 2000.

[8] Sistrunk, W. "Dating LPs," *Music Reference Services Quarterly,* 8, 4 (2004), 47–55

[9] Mudge, S., and Hoek, D. "Describing jazz, blues, and popular 78 rpm sound recordings: Guidelines and suggestions," *Cataloging & Classification Quarterly,* 29, 3 (2000), 21–48.

[10] Simpkins, T. "Cataloging popular music recordings," *Cataloging and Classification Quarterly, 31,* 2 (2001), 1–35.

[11] Smiraglia, R. *Describing music materials,* 3d ed. Soldier Creek Press, Lake Crystal, MN, 1997.

[12] Hemmasi, H. "Why not MARC?" *Proceedings of the International Conference on Music Information Retrieval,* 2002.

[13] Categories for the Description of Works of Art. Retrieved July 23, 2005 from http://www.getty.edu/research/conducting_research/standards/cdwa/

[14] Visual Resources Association Data Standards Committee. VRA Core Categories, Version 3.0. Retrieved July 23, 2005 from http://www.vraweb.org/vracore3.htm

[15] Harvard University Library. "Administrative metadata for digital still images, version 1.3", Harvard University Library Specification, 2004. Retrieved July 23, 2005 from http://preserve.harvard.edu/resources/imagemetadata.pdf

[16] National Information Standards Organization. "Data dictionary: Technical metadata for digital still images," Working draft, 2002. Retrieved July 23, 2005 from http://www.niso.org/pdfs/DataDict.pdf

[17] Hurley, B., Price-Wilkin, J., Proffitt, M., and Besser, H. *The making of America II testbed project: A digital library service model.* The Digital Library Federation, 1999

[18] Library of Congress. Table of core metadata for LC digital repository development, 2000. Retrieved July 23, 2005 from http://www.loc.gov/standards/metable.html

[19] Koenen, R., and Pereira, F. "MPEG-7: A standardised description of audiovisual content," *Signal Processing: Image Communication,* 16, 1 (Sept. 2000), 5–13.

[20] California Digital Library. "Digital image format standards," *CDL Reports & Guidelines*, 2001.

[21] Bradley, K. ed. *Guidelines on the production and preservation of digital audio objects,* Aarhus, Denmark: International Association of Sound and Audiovisual, 2004.

[22] Ester, M. *Digital image collections: Issues and practices*. Commission on preservation and access, Washington DC, 1996.

[23] Yorkston, S., and Quantum-X. "What does the 'Quality' setting mean?" Retrieved July 23, 2005 from http://www.vorbis.com/faq.psp#quality

[24] Droettboom, M., MacMillan, K., and Fujinaga, I. "The Gamera framework for building custom recognition systems," *Proceedings of the Symposium on Document Image Understanding Technologies,* 2003.