Archiving David Edelberg's Handel LP Collection: Production Workflow and Issues in Data Acquisition

Catherine Lai and Ichiro Fujinaga; Schulich School of Music, McGill University, Montreal, Quebec, Canada;

Abstract

This paper describes the production workflow for building a digital archive of David Edelberg's Handel LP collection and discusses the issues in phonograph data acquisition in general. Metadata elements from the metadata data dictionary for phonograph records are listed.

Introduction

Libraries worldwide are realizing the potential of digital information and initiating the digitization of their rare and unique cultural materials. One of the most precious holdings of McGill University's Marvin Duchow Music Library is the irreplaceable analogue sound recordings of Handel's music assembled by David Edelberg (1939–1989). David Edelberg started assembling Handel long-playing phonographs (LPs) in the mid-1970s. Over the subsequent fifteen years, his collection grew to nearly 3,000 sound recordings [1]. Among these analogue recordings, more than 130 of the LPs are now in the public domain as they are recordings of Classical music released before 1956. Approximately thirty LPs from the collection have been digitized as a pilot study. Digitizing these LPs is beneficial to the collection as it preserves valuable cultural heritage material and provides wider access to the recordings. Furthermore, the experience gained in this project will hopefully be useful in other similar projects.

Production Workflow

To assure the quality of preservation and to facilitate new forms of use and access, an extensive digitization workflow management system has been developed. This workflow involves hardware acquisition, installation, and configuration; software installation and development; copyright and rights management; metadata scheme design; analogue-to-digital (A/D) audio conversion; audio track separation; image scanning of record labels and packaging (album covers and liner notes); metadata extraction; text conversion; creation of derivatives; database design, implementation, and maintenance; web site design, development, and maintenance; and review and evaluation of the system.

The production workflow of the Edelberg Handel collection is unique for several reasons. The archiving project develops an efficient and economical framework to manage large-scale digitization of musical material—LPs, and the production involves digitization of both audio and visual components (i.e., album covers and liner notes). Furthermore, the workflow management system provides benchmarking for conversion and access and creating ground-truth data to train and test content analysis systems.

Control the Quality Control Environment

The quality of digital reproduction, to a significant degree, relies on the quality control of the device and software used in the chain of digitization [2]. The Handel digitization project uses state-of-the-art digitization equipment and software tools. 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. Software tools used relating to image-display conditions include color management, which ensures color consistency from image capture through display, and monitor calibration/optimization, which ensures onscreen accuracy by setting white and black point, gamma, and color balance.

Copyright and Rights Management

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 and renewed or extended copyrights due to re-release of album records can occur, legal clearances of LPs for digitization required complex rights management checks via various sources such as the WorldCat OCLC Union Catalog, the Bielefelder Catalog, the Diapason Catalog, the Schwann Catalog, and the Gramophone Catalog. In addition to finding out the album release date to determine the legal status of the LPs, verifying the status of all the copyrighted materials (e.g., photographs, program notes) that appear as an integral part of the album were also necessary.

Digitization

An LP must be as clean as possible to achieve optimum audio quality. Therefore, each side of the records was thoroughly vacuum cleaned with record cleaning solution before each audio digitization session to remove accumulation of dirt that may consist of grease, dust, or surface particles. The audio was digitized at 24bit/96kHz and saved in the industry standard AIFF format. The cleaning and digitization of one side of audio disc took approximately 30 minutes. 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 in lossless compression PNG format. To ensure color consistency across a wide range of display equipment, a small color separation guide (Kodak No. Q-13) composed of a set of standard color patches (primaries, white, and black) was always placed in the same position relative to the scanned material. Scanning an image at 24bit/1200dpi took approximately 13 minutes and saving the file to disk took an additional 12 minutes.

Content Management

A web data-entry form was implemented in PHP for data and metadata entry. The entry form used 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. The form, moreover, 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.

Database Design

A relational database in MySQL was designed and implemented to hold the metadata of the digitized material. Data verification and database tuning were conducted iteratively throughout the digitization process for quality assurance and performance considerations.

Metadata Extraction and Text Conversion

Metadata extraction and text conversion required much human intervention and a high-level musical and bibliographic knowledge. Information such as the location of diverse material on the album covers (e.g., phonographs) and the text column and typography of the text of the printed material was meticulously recorded to be used later as ground-truth data for developing automatic document analysis software using Gamera [3]. Another expensive step of digitization is text conversion of program notes on album covers or any accompanying material. Text conversion required typing columns of text that took at least 20 minutes per column of text of approximately 670 words. An average of six hours was needed to process a phonograph album during the pilot study. Although taking the physical measurement was extremely time consuming, even without this requirement, the process would still take about three hours per phonograph album.

Creation of Derivatives

Image and audio formats were optimized for different purposes. Different technical specification were examined and developed for the derivative files for web access and print versions. Although there are several standards governing the creation and use of digital images, such as the California Digital Library's format standards [4] and the IASA Guide [5], the unique characteristics of the master files require different format specification for different purposes [2]. Evaluation criteria for technical specification for image derivatives included attention to the legibility of the smallest text, preservation of color appearance, and speed of delivery.

Five types of access files were generated in JPEG: A print file (300 dpi), two versions of web display files (96/120 dpi), and a thumbnail file (72 dpi) with its popup image (600x600 pixels). A resolution at 96 dpi was chosen for the web access file of album cover images and a resolution at 120 dpi was chosen for the web access file of label images because the disc labels often contain very small texts that require higher resolution to ensure legibility of the smallest text.

Access or derivative audio files were created in various levels of fidelity in different formats: MP3 files of 112kbps and 192kbps, WAV files of 16bit/44.1kHz, and Ogg Vorbis files of quality 5. Access files in MP3 format were created in higher and lower fidelity to provide good-quality audio with smaller file sizes. Wav files were created to offer high-quality audio that can be easily used on both Macs and PCs. Ogg Vorbis derivative files were created because the format is free, open and unpatented. Furthermore, Ogg quality 5 is the official setting recommended for representing music of CD quality [6].

Web Delivery

A web site implemented in PHP facilitates easy access to the digitized recordings of David Edelberg's Handel LPs. The digital archive is accessible by browsing or searching on any word in the metadata database. Brief records appear in search results and display summaries of the records, including the collection number, album title, series information, and label issue number. Full records feature the complete record, including access to image and audio files.

This site provides online access to the intellectual content as well as reproductions of both images and audio. For each record, it displays all the metadata associated with an LP, links to separated and continuous audio tracks as well as scanned images of album covers and any accompanying material. Due to variations in computer platform and connection bandwidth, multimedia files in different formats and resolutions are offered to meet diverse user needs. The website is hosted on a RAID 5 server and is mirrored on a duplicate system in a separate building for redundancy and backup.

Issues in Data Acquisition

Although traditional standards for cataloguing sound recordings exist [7, 8, 9], these formats are inadequate for describing and searching digitized representations of LPs. The current practices are generally limited to bibliographic description of relatively few elements [10]. Information about artwork or photographs in the album packaging of a LP, for example, is usually not included. A few recognized authorities have begun to contribute and expand the utility of metadata to sounds. MPEG-7, a formal system for describing multimedia content, defines elements for description of audio and video content [11]. However, the schema does not have characteristics tailored to the structural complexity that is necessary for the full description of sound recordings. For example, significant types of important information at the individual track or song level are missing. Unspecified information such as recording location, recording session date, recording engineers, or recording equipment used are potentially useful and convenient access points.

To better facilitate the management and use of resources from the intellectual content to the technical and legal information, including provenance, authenticity, and intellectual property and rights metadata, a data dictionary—a formal specification of data definition—has been developed. The data dictionary clarifies the scope and types of metadata associated with digital objects, which helps to prevent duplicate handling of data, inconsistencies, and lack of integrity during data entry as results of interpreting at different levels of summarization. For examples, the dictionary makes clear distinctions between album title and musical work title and total duration and track duration. The data dictionary also provides guidance for data entry decision-making: it gives formal

definition, usage notes, and data type to each metadata. The technical specification disambiguates metadata terms (e.g., label issue number vs. matrix number), thereby clarifying the conceptual intent. Moreover, the data dictionary reinforces vocabulary control for data values. Photos and artwork on album covers, which reflect social context, are being documented and described using control vocabularies to enhance access to information on the visual components of the LPs. The complete set of metadata elements describing phonograph records is included in Appendix A.

Future Work

For future work we plan to digitize the remaining LPs now in the public domain. Due to the enormous quantity of existing recordings and the time required to properly and faithfully digitize them, an important next step in this digitization research project is to build upon the ground-truth data already captured by integrating sophisticated pattern recognition systems to automatically generate text and metadata from the captured images.

Conclusion

Currently there is no publicly accessible large-scale classical music audio file depository anywhere in the world. The completion of the digital collection of Edelberg's LPs, especially with the searchable metadata and full textual searching of content, provides a digital archive that is unique, groundbreaking, and significant. As librarians and archivists continue to recognize the need to digitize their analogue sound recording collections, the need for efficient workflow management tools increases. The methodology and tools developed here are to be made available to other libraries and archives, hopefully promoting similar digital archiving projects.

Acknowledgement

This research is funded in part by the "Richard M. Tomlinson Digital Library Innovation and Access Award," David Edelberg Foundation, CIRMMT, CFI, and FQRSC.

References

[1] Marvin Duchow Music Library, David Edelberg. (2004). Retrieved March 5, 2006. Available at http://music.library.mcgill.ca/edel_fr.htm.

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

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

[4] California Digital Library, "Digital image format standards." CDL Reports & Guidelines. (2001).

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

[6] Xiph.Org., (1994). Retrieved March 5, 2006. Available at http://www.vorbis.com/faq/

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

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

[9] R. Smiraglia, *Describing music materials,* 3rd ed. (Soldier Creek Press, lake Crystal, MN, 1997).

[10] H. Hemmasi, "Why not MARC? " *Proceedings of the International Conference on Music Information Retrieval*. pg. 242–8. (2002).

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

Author Biography

Catherine Lai received her BAs in music and applied math with an emphasis in computer science and M.I.M.S. in information management and systems from the University of California at Berkeley. She is currently a Ph.D. student in Music Technology at McGill University. The focus of her research is on music information retrieval. She is interested in exploring web-based music information retrieval constituting areas such as data/information presentation and user interface design.

Ichiro Fujinaga is Assistant Professor of Music Technology at McGill University. His research interests include optical music recognition, music perception, machine learning, and music information retrieval.

Appendix A

The complete set of metadata elements describing phonograph records is listed below by type:

Description metadata (collection level)

- Collection ID
- Summary of Collection
- Subject of Collection
- External Review of Collection
- Scope of Collection
- Content of Collection
- Type of Collection
- Holding Institution of Collection

Description metadata (album level)

- Title of Album
- Varying Form of Title
- Uniform Title
- Language of Album Text
- Label Name
- Label Issue Number
- Cross Label Name and Issue Number
- Other Catalog Number
- Re-issue Date
- Series Statement
- Volume Number
- Edition Statement
- Total Duration
- Date/Time of Event
- Form of Musical Composition
- Price Information
- Number of Accompanying Material
- Accompanying Material Characteristic
- Physical Dimension of Accompanying Material
- Number of Audio Discs
- Composer Information
- Performer Information
- Instrument Group Information
- Recording Engineer Information
- Lyric Writer Information
- Arrangement Writer Information
- Chorus Information
- Artwork Information
- Recording Equipment
- Date of Recording
- Place of Recording
- Recording Engineering Detail
- Name of Manufacturer
- Manufacturer Information
- Name of Distributor
- Distributor Information
- Advertisement
- Warning
- Handling Instructions
- Reviews
- Acknowledgement
- Disclaimer
- Music Work Notes
- Artist Notes
- Libretto Notes
- Other Notes
- Language of Accompanying Material
- Writer Name
- Written Date
- Writer Birth Date
- Writer Death Date
- Writer Flourished Date

Description metadata (artwork level)

- Artwork Description
- Function of Artwork
- Caption of Artwork
- Date of Artwork
- Artwork Artist Birth Date
- Artwork Artist Death Date
- Artwork Artist Flourished Dates

Description metadata (audio level)

- Audio Disc Number
- Side Number on Audio Disc
- Number of Tracks
- Matrix Number
- Audio Disc Dimension
- Type of Recording
- Playing Speed
- Groove Characteristics
- Number of Sound Channel
- Audio Disc Notes
- Audio Disc Peculiarity Notes

Description metadata (track level)

- Track Number
- Title of Work for Track
- Part Name of Work
- Work Catalog Number
- Key of Music
- Language of Work
- Medium of Performance
- Arrangement Statement
- Date of Work/Composition
- Track Duration
- Date of Recording
- Place of Recording
- Recording Engineering Detail
- Track Notes
- Track Peculiarity Notes
- Instrument Group
- Artist Name
- Artist's Instrument/Role
- Artist Birth Date
- Artist Death Date
- Artist Flourished Dates

Administration metadata

- Source Item ID
- Source Type
- Provenance of Data
- Comment about the Provenance of Data

Structure metadata

- Typography of Data
- Font Size of Data
- Layout of Data

Legal rights metadata

- Trademark Registration Information
- Copyright Registration Number
- Copyright Assignment Number
- Copyright Display Text
- Copyright Begin Date
- Copyright End Date
- Patent Registration Information
- Patent Assignment Number
- License/Patent Display Text
- License/Patent Begin Date
- License/Patent End Date

Technical information metadata

- Digitization Project Name
- Digitization Project Number
- Digitization Funding Information
- Digitization Source Item ID
- Digitization Source Type
- Digitization Source Characteristics
- Electronic Access Location
- Date of Digitization
- Place of Digitization
- Transcriber
- Producer
- Digitization Notes
- Sequential Production Notes

Technical information metadata (image)

- Scanning Hardware
- Display Equipment
- Physical Dimension of Source
- Physical Dimension of Area Scanned
- Light Source
- Color Management
- Color Bit Depth
- Image Resolution
- Master Image File Format
- Image Compression
- Control Target
- Color Bar
- Image Software
- Image Editing

Technical information metadata (audio)

- Audio Capture Device
- Stylus Dimension
- Stylus Shape
- Stylus Tip Mass
- Tracking Force
- Tonearm and Cartridge Alignment
- Anti-skate
- Turnover
- Rollover
- Playback Speed
- Record Cleaning
- Stylus Cleaning
- Record Repairs
- Audio Bit Rate
- Audio Resolution
- Number of Recording Channel
- Recording Level
- Digital Noise Reduction
- Master Audio File Format
- Other Capture Detail
- Audio Software
- Digital Audio Editing