# The Challenges in Developing Digital Collections of Phonograph Records

Catherine Lai
Faculty of Music
McGill University
Montreal, QC H3A 1E3
+1 514 398-4535 x0300
lai@music.mcgill.ca

Ichiro Fujinaga
Faculty of Music
McGill University
Montreal, QC H3A 1E3
+1 514 398-4535 x00944
ich@music.mcgill.ca

Cynthia Leive
Marvin Duchow Music Library
McGill University
Montreal, QC H3A 1B9
+1 514 398-4694
cynthia.leive@mcgill.ca

## ABSTRACT

To facilitate long-term preservation and sustain the utility of phonograph records, an efficient and economical workflow management system for digitization is necessary. We describe in this paper the digitization process for building an online digital collection of phonograph records and our procedure for creating the ground-truth data, which is essential for developing an efficient metadata and content capturing system. We also discuss the challenges of defining metadata for phonograph records and their packaging to enhance access and use across traditional boundaries.

## Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Collection and Standards.

## General Terms

Management, Standardization.

## Keywords

Digitization, Preservation, Metadata, Analog Sound Recordings, Phonograph Records, Use and Access, Digital Library Collections, Music Information Acquisition and Retrieval.

## 1. INTRODUCTION

Long-playing phonograph records (LPs) were distributed commercially throughout most of the twentieth century. Many of these historic sound recordings have long shelf lives; however, several compelling reasons have led to a shift toward digital preservation. First, LPs are at risk of deterioration because phonograph discs are fragile and ephemeral as their physical composition does deteriorate upon playback. Second, many rare and important recordings are being discarded because the appropriate playback equipment is becoming scarce and cumbersome to maintain. Also, although traditional preservation methods in libraries and archive communities have ensured the longevity of these endangered materials, storing and protecting against risks of damage or misuse in archival custody have sometimes been at the cost of reduced access. Since in many countries, including Canada, recordings of Classical music released before 1955 are now in the public domain, digitizing these recordings and their album covers and making them freely available on the Internet will be extremely valuable to musicological research and music education and provide a valuable source of enjoyment for the general public.

To provide preventative preservation and new forms of access to these very important cultural heritage materials, a large digitization effort is required. An efficient and economical workflow management system is essential to carry out the steps in the digitization process. Major tasks of digitizing LPs include analog-to-digital audio conversion, audio track separation, image scanning of record labels and packaging (album covers and liner notes), metadata extraction, and text conversion. This process is time-consuming and expensive since the steps in digitization require much human intervention and a high-level musical and bibliographic knowledge. Thus to digitize millions of LPs that exist in a reasonable amount time, it is important that the process of digitization be as efficient as possible.

To minimize human intervention, thereby reducing the cost of digitization, we propose to integrate sophisticated pattern recognition systems to automatically generate text and metadata from the captured images. One strategy is to develop a document-understanding software that will recognize the structure of the album covers and segment the scanned images into photographs, graphic arts, columns of text, etc. The results of which can be used as sources for optical character recognition, picture analysis, or metadata extraction systems. Much of the effort at this initial stage of the current project was devoted to creating ground-truth data that can be used to train and test the system. Another time-consuming task, if performed by dedicated human digitization operators, is separating music tracks that exist on each side of discs. We are developing software to automatically separate the tracks knowing the number of tracks, which can be extracted from the record label or the album cover. In order to improve this software, we also require ground-truth data of correctly separated audio tracks.

As a pilot digitization project of phonograph recordings, we digitized approximately twenty LPs from David Edelberg's Handel collection, which is one of the largest collections of analog recordings of Handel's music, housed in McGill Music Library.

## 2. BACKGROUND

Developing digital collections of LPs presents new challenges. Although recent technology and software, for examples, Olive

*JCDL'05*, June 7–11, 2005, Denver, Colorado, USA

Software, TEXTML and Adobe XMP, have been developed and employed by newspaper organizations, magazine companies, and libraries to bridge the gap between the print, archive, and online accessibility, the structural complexity of LP and its packaging demands more than their content management models can offer. In addition to basic music attributes such as song titles or composer names, an LP record album can also include additional information such as album cover picture, translation of opera text, track credits, matrix numbers, and more, which should all be easily accessible and searchable.

# 3. WORKFLOW MANAGEMENT

A workflow management system for digitizing LPs was developed and involved several steps including metadata extraction, analog-to-digital audio conversion, image scanning, text entry, and creation of derivatives of audio and images for online delivery. In order to create a flexible and relevant archive, a new metadata schema for sound recordings was developed.

## 3.1 Metadata Schema

A comprehensive metadata schema for describing LPs has been designed to the finest level of granularity possible as part of the large digitization management system. The schema includes five types of metadata: description (enable discovery and identification of resources), administration (support management of resource), structure (describe font and layout characteristics of texts), legal rights (protect intellectual property rights), and technical information (record the capture process and technical characteristics of the digital objects). The new metadata schema provides for complete auditory, pictorial, and textual content analysis. Characteristics from Dublin Core, MARC21, MODS, METS, TEI, MPEG-7, and more were partially incorporated into its design. The current schema contains more than 120 fields.

## 3.2 Metadata Extraction

A web data-entry form written in PHP was implemented for the encoding of the data and metadata of the LPs. The metadata entry system enforced quality control, using check boxes and option buttons whenever possible to reduce typing errors. The form also employed some error checking, validating data before submitting them into a relational database (MySQL) designed and implemented to hold the metadata and the content of the digitized material. Information such as the location of diverse material on the album covers (e.g. photographs) and the text column and typography of the text of the printed material was recorded to be used later as ground-truth data for developing automatic document analysis software using Gamera [1].

The manual entry in this initial experiment took an average of six hours to process a phonograph album. This included audio digitization; scanning of album covers, liner notes, and discs; metadata extraction, which included measuring physical positions and size of the visual contents of the album; and full-text entry. Although taking the physical measurement was extremely time consuming, it is estimated that even without this requirement, the process will still take about three hours per phonograph record album.

## 3.3 Digitization

The digitization process consisted of vacuum cleaning each side of the discs; digitizing audio at 24bit/96kHz using an audiophile-quality turntable and a cartridge; and scanning images, including the album covers, audio discs (for labels and matrix numbers), and any accompanying materials, at 24bit/1200dpi. The aim is to produce digital audio and images for an enduring and high-quality archive.

## 3.4 Creation of Derivatives

With the exception of manually separating audio tracks, the process of creating different versions of audio and images from master files was accomplished automatically using open-source software: sndfile-convert, SoX, LAME, and ImageMagic. Derivatives of audio and image files provide online presentation and accessibility of data in various resolutions and formats.

# 4. CHALLENGES

The structural complexity of the music and LPs imposes many challenges to developing digital collections of LPs. One of the challenges in defining metadata for LPs is to determine the level of detail and maintain various kinds of metadata. In addition, to define the *content* of the data, which is what the object contains or is about (intrinsic), decisions also had to be made on how to define the *context* of the data (extrinsic, e.g., font properties and text properties) as well as the *structure* of the data (intrinsic or extrinsic, e.g., positioning context in relations to the other block elements) [2]. 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. Another major challenge is how to develop and adopt an automatic name authority control to manage variations in spelling, e.g., names (composers, performers, producers) and musical work titles. Other challenges include content description for images on the album covers and the complex rights management for different elements of LPs, including photograph, artwork, trademarks, music, music arrangements, lyric, etc.

# 5. FUTURE WORK AND SIGNIFICANCE

An immediate task for this project is to separate automatically music tracks using digital signal classification techniques. Another more challenging future work is to automate the metadata and content extraction. We plan to achieve this goal by using the ground-truth data already captured in this project and further developing Gamera. Developing digital collections of analog holdings such as LPs, libraries will enhance access and use of valuable but rare analog sound recordings across traditional boundaries, extend the reach of research and education, and ensure the persistence of LPs via preventative digital preservation as part of the cultural necessity.

# 6. ACKNOWLEDGEMENTS

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

# 7. REFERENCES

[1] 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.

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