Case study

Transitioning the largest archive of animal sounds from analogue to digital

Received (in revised form): 8th October, 2013

Gregory F. Budney

is Curator of Audio for the Macaulay Library at the Cornell Lab of Ornithology. He oversees the Macaulay Library audio preservation efforts and has had the pleasure of serving archive since 1980.

Macaulay Library, Cornell Lab of Ornithology, Ithaca, NY 14850, USA
Tel: +1 607 254 2406; E-mail: gfb3@cornell.edu

William McQuay

is the Supervising Audio Engineer for the Macaulay Library at the Cornell Lab of Ornithology. Prior to joining the Macaulay Library in 2007, he was technical director of National Public Radio's 'Radio Expeditions' series and the classical music show 'Performance Today'.

Macaulay Library, Cornell Lab of Ornithology, Ithaca, NY 14850, USA
Tel: +1 607 254 1137; E-mail: wjm93@cornell.edu

Michael S. Webster

is Director of the Macaulay Library and the Robert G. Engel Professor of Ornithology in the Department of Neurobiology and Behaviour, Cornell University. In addition to overseeing all activities at the Macaulay Library, he conducts research on animal communication signals, particularly in birds.

Macaulay Library, Cornell Lab of Ornithology, Ithaca, NY 14850, USA
Tel: +1 607 254 2493; E-mail: msw244@cornell.edu

Abstract The Macaulay Library is the world's largest scientific collection of audio and video natural history recordings. This paper focuses on a recently completed multi-year effort to digitise the archive's collection of analogue-archived audio recordings. The paper describes that effort, the main principles that guided it and the lessons learned. In particular it emphasises the need to preserve audio assets at the highest possible technical standards, to use technologies with wide industry support and to use a format that facilitates migration to future formats and storage systems. The importance of rich metadata and accessibility is also discussed. The effort to digitise this collection opened it to the world, leading to dramatic increases in its use for diverse purposes ranging from scientific research to the arts.

KEYWORDS: audio, digitisation, natural history recordings

INTRODUCTION

The Macaulay Library (ML) at the Cornell Lab of Ornithology (birds.cornell.edu) is the world's largest and oldest archive of natural history recordings. The library's holdings are vast, with currently over 195,000 audio and 60,000 video assets that capture the communication signals and other behaviours of over 9,000 species of animals; this includes not only 7,000 species of birds, but also a diversity of mammals, reptiles, amphibians, fish, insects and even environmental recordings. As such, the ML is not only a valuable resource for biologists and other research scientists, but also for nature enthusiasts (eg bird watchers), filmmakers, musicians and other artists. Because the roots of the archive stretch back to the early 20th century, the majority of these recordings were in an analogue format, primarily magnetic tape, which made them relatively inaccessible and created challenges for those wishing to find recordings for their work, be it a research project or a musical composition. Accordingly, an important step for bringing the archive into the 21st century was to digitise the fully archived assets on analogue reel-to-reel tape and make them broadly accessible. The effort to do this was completed in October 2012 and today over 160,000 audio and 50,000 video recordings are digitised and playable online at the ML website (Macaulay Library.org). This paper briefly outlines the steps and main decisions that were made during this digitisation process and also the steps that are currently being undertaken to make the digital archive even more accessible and useful for future generations. The focus is primarily on the digitisation of the audio collection, although most of the general points are also applicable to video and other media types.

HISTORY: WHERE IT STARTED

The origin of the ML traces back to the very first audio recordings of wild birds ever made in North America. In May 1929, film pioneer Theodore Case sought the assistance of Cornell University professor of ornithology Arthur A. Allen to record the songs of wild birds as a means of demonstrating the potential of a system for synchronising sound with motion picture film (ie 'talkies'). This was the birth of the Library of Natural Sounds (LNS), later to become the ML. Today, more than 80 years later, one may listen to two of the three birds recorded on that morning in May: a rose-breasted grosbeak (http://macaulaylibrary.org/audio/16968/) and a song sparrow (http://macaulay library.org/audio/16737/), via the ML website. Although the LNS was housed at the Lab of Ornithology and started with recordings of birds, it quickly grew to include recordings of other types of animals as well. Indeed, those early nature recordists quickly discovered that recording amphibians such as frogs and toads had its advantages, as the sounds they made were complex and the quarry did not fly away when approached. Since those earliest recordings of birds and frogs, the ML (then LNS) has been dedicated to recording, preserving and disseminating wildlife recordings for a diversity of uses, and the recordings themselves have been treated as irreplaceable biological data or 'media specimens'.

Not surprisingly, given its long history, the analogue media used to collect natural history recordings in the field and store them at the ML have evolved over the years (Figure 1). Throughout the 1930s and into the early 1940s, sound-synch film was the medium used by Allen and his cohorts, Peter Paul Kellogg and Albert R. Brand. The year 1944 brought a transition to cutting acetate records in the field and with it came the near-instant ability to verify the results of recording efforts. In

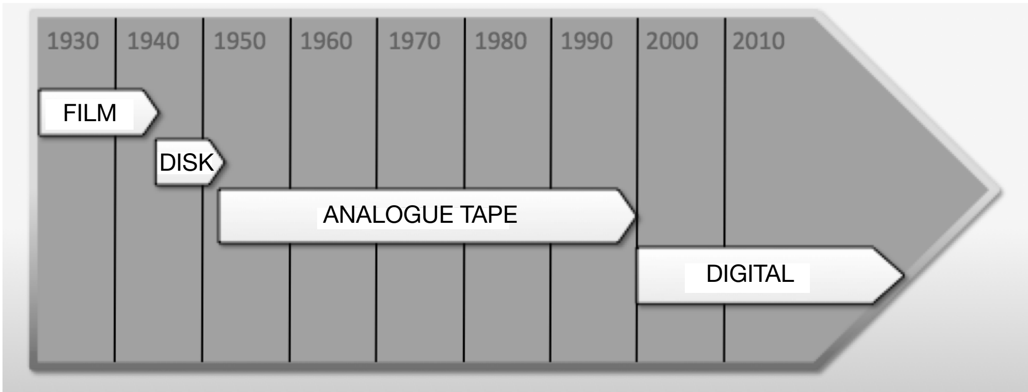

Figure 1: The Macaulay Library: 80 years of archival format evolution

the 1950s, with the advent of magnetic tape and reel-to-reel recorders, both field recording and storage of audio assets migrated to acetate-based tape. To facilitate curation and retrieval, field tapes were painstakingly cut and spliced together to produce 'species reels' — reels that contained the recordings of just a single species; requests for material were fulfilled by making copies from these spliced species reels. At the same time, due to the safety hazard posed by cellulose nitrite film recordings, valuable assets recorded earlier were transferred either to disk or reel-to-reel, with the source nitrite film subsequently destroyed. The archival protocol then underwent an important modification, original field recordings were duplicated at 15 inches per second (ips) and those mylar-based copies were used to construct species reels and serve as the working copy. Aside from a shift away from acetate-based media to mylar tape stocks, and also the addition of analogue cassette (1970s) and rotary digital audio tape (1990s) as field recording formats, the archival process remained fundamentally unchanged for over 50 years. It is a tribute to the format's robustness that, today, it is still possible to accurately play back a 50-year old analogue reel-to-reel recording. Thus, by the end of the 20th century, the ML holdings stood at 130,000 audio recordings, stored primarily on analogue reel-to-reel tapes organised by species.

Up to this time, despite being technically feasible, ML staff had resisted encouragement to convert the archive from analogue to digital CD-based storage. This was primarily because the 44kHz/16-bit audio standard, which was used for CDs and widely available at the time, was considered by the library staff to be inadequate for archival of many wildlife sounds. Similarly, the broader audio archival community recommend the use of digital sampling rates higher than 44.1 kHz,[1] as higher sampling rates yield superior temporal resolution for detailed time-based analysis and greater precision for localising sounds (with stereo or surround sound recordings). This also enabled manufacturers to build better anti-aliasing low pass filters that operate more efficiently, thereby improving system performance. Conversely, despite its high audio fidelity and longevity, the analogue tape format had a limited lifespan as tapes inevitably degrade and completely accurate tape duplication remains impossible. Moreover, the analogue tape format posed a serious impediment to servicing the requests from the archive's clients, which ranged from scientists to the general public, to museums, to aquaria and to film production companies. Requests for modest numbers of recordings could take weeks to fulfil and requests requiring copies of hundreds of recordings could take months. A paradigm shift was needed and it came in the form of the industry acceptance of the DVD-Audio (DVD-A) standard, which allowed the use of 96 kHz/24-bit high-resolution recording.

INTO THE DIGITAL AGE

From the onset of digitisation up to today, strategic bases of the efforts to digitise the collection have been to preserve audio assets at the highest possible technical standards, to use technologies with wide industry support and to use a format that facilitates migration to future industry-supported storage strategies and systems. The 96 kHz/24-bit standard fulfilled all three of these criteria and is now recognised as the standard for audio preservation reformatting.[2] The time to 'go digital' had arrived and so, in early 2000, with major support from the National Science Foundation and later also from the Office of Naval Research (for digital archival of marine recordings of mammals and fish), the effort to digitise the entire analogue audio archive was begun.

Due to its excellent high-resolution fidelity and potential longevity, DVD-A was adopted as the format for the digitised audio collection. Although DVDs provided excellent long-term archival storage, they nonetheless created challenges, as the format did not allow for easy sound file retrieval. The initial approach to addressing this challenge was to install two Plasmon DVD 'jukeboxes', each capable of housing 480 disks and with file retrieval capability of approximately seven seconds. Although a step forward, this DVD-based system could not serve the demands of online access and rapid internet (eg via ftp) dissemination. A second challenge was file migration, as DVD-A had an uncertain pathway for future file migration and digital file migration is critical to meeting the inevitable changes in digital format and technology that would come in the future.[1] Finally, DVD-A was a proprietary file format requiring specialised equipment for playback. To address these various challenges, in May of 2001, ML acquired its first storage area network, a 7TB EMC2 RAID 10 SAN, and the migration of high-resolution audio files from DVD to enterprise hard-drive storage commenced. From this point forward, each specimen was archived as a high-resolution (96 kHz/24-bit) file (originally an AIFF file, later replaced by wav and then bwav files, as these are non-proprietary formats that do not require specialised hardware or software to read). Lower resolution derivative files (44kHz/16-bit and mp3) were also made and stored for more rapid web-based delivery. Although the working archive was stored on hard drives, DVD storage was still regarded as the primary 'deep' storage because a fixed medium (like DVD) was considered to have a better shelf life than spinning disks and a redundant array of independent disks (RAID) was not considered adequate for long-term preservation.[1]

The effects of the new digitisation process were immediate. Within one year, the amount of time required for an audio archivist to process one hour of audio and associated metadata was reduced from eight to ten hours to two to three hours through the elimination of entire procedural steps, not the least of which was the time-consuming process of splicing of 15 ips copies onto archival reels organised by species along with the associated requisite labelling and cataloguing procedures of these physical media. Moreover, with the archive digitised, it was possible for the public to

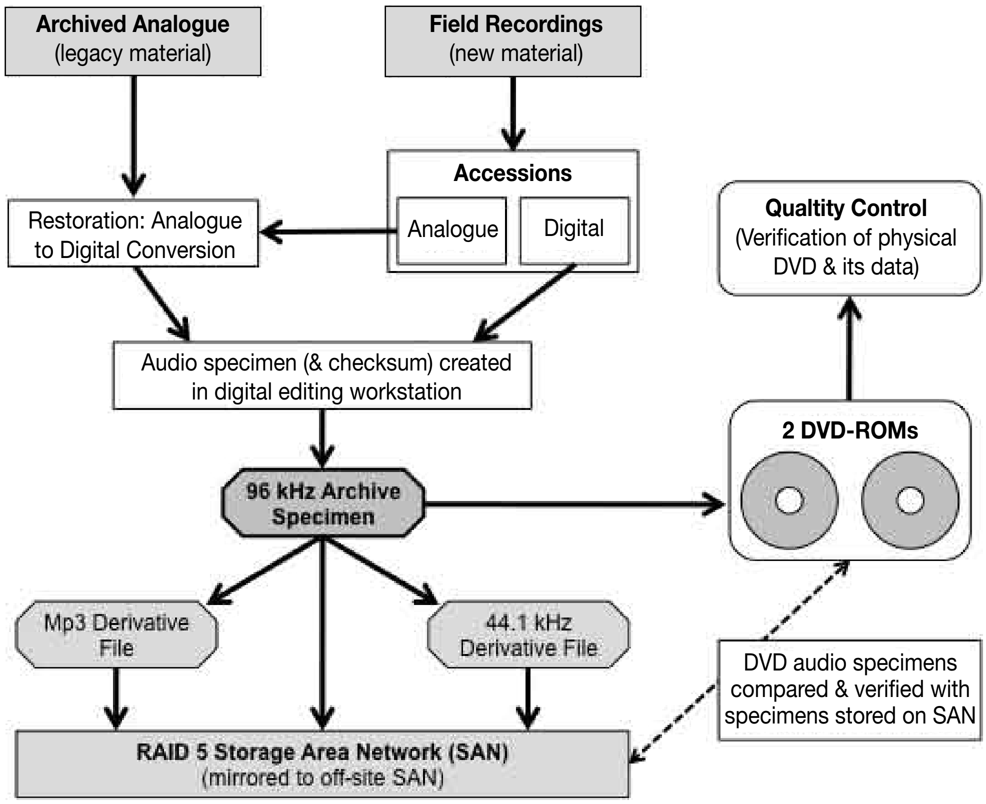

Figure 2: The Macaulay Library digital audio workflow

explore it and listen to low-resolution (mp3) files online, and on demand, at MacaulayLibrary.org. Thus, the 12-year effort to digitise the ML audio collection made over 7,000 hours (or 313 days) of audio specimens accessible and thereby opened the doors of the entire archive to the world.

DIGITISATION, STORAGE AND QUALITY CONTROL: SOME DETAILS

A media archive like the ML, which is interested in preserving recordings from the past as well as those made in the future, faces a unique challenge. While it is critically important to facilitate the ability of recordists to collect and archive recordings collected on modern-day digital equipment, it is simultaneously important to be able to take in analogue recordings collected using technologies from the past. This is a situation that will persist for a number of years as researchers and private recordists, whose active work period spanned the latter half of the 20th century, seek a repository to preserve their work. Accordingly, it is critically important to maintain legacy analogue playback equipment, ensuring properly calibrated performance. Trained professionals capable of using sophisticated test equipment and established testing modalities are therefore essential personnel. In addition, it is necessary to have trained staff capable of running the equipment to digitise the recordings. Ideally, these personnel should be knowledgeable in restoration techniques required to address a plethora of problems associated with ageing magnetic tape.[2] Toward these ends, the ML employs a staff of media archivists and media engineers trained in the appropriate methods and technologies required to

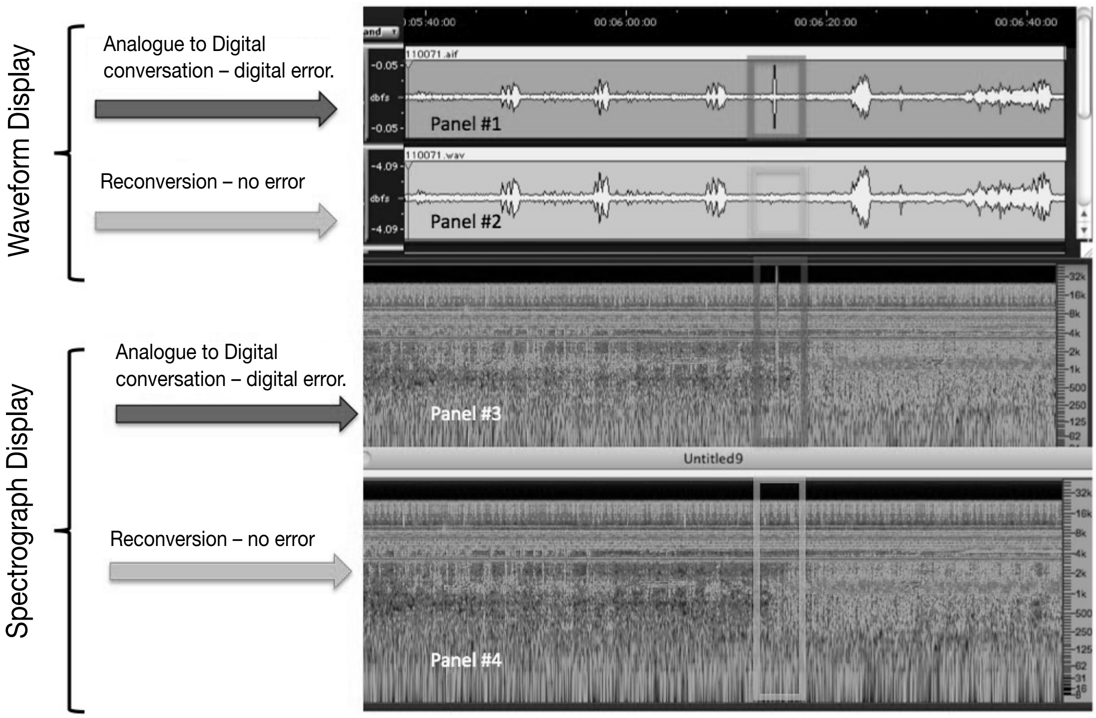

Figure 3: Waveform, spectrogram and oscilloscope verification of A/D conversion. A visual display of an analogue to digital conversion used during QC reveals a possible digital error (panels 1 and 3). This is confirmed when the sound source is reconverted (panels 2 and 4.) An investigation revealed that the A/D converter may have been improperly set.

accomplish these tasks to established standards.[3]

The ML digital archive is based on the Open Archival Information System model, which is a widely adopted conceptual model for a digital repository and archival system.[1] Digitisation workflow involves a number of important steps (Figure 2). In general, a ML audio archivist first performs an assessment of the condition of the analogue media to determine whether physical restoration is required and, if so, performs the appropriate restoration action. After tape restoration, the archivist works in the studio to convert the analogue signal to digital. Each of the ML audio studios contains a digital audio workstation capable of creating high-resolution audio DVD masters, precision analogue/digital converters with extremely accurate digital clocking and discrete working storage. The system permits an archivist to digitise the analogue material and compare playback of the source against downstream digital files via near-field audio speakers as well as waveform, spectrogram and oscilloscopic displays to verify the integrity of the conversion process (Figure 3). Ahead of work by audio archivists, audio engineers are responsible for ensuring the performance of ingest and archival systems, regularly reviewing parameters such as equipment signal:noise ratio, channelisation, spectrum analysis of the noise floor and harmonic distortion and total harmonic distortion plus noise versus input frequency response. These quality assurance efforts identify and correct equipment issues as soon as practically possible, so work proceeds with confidence from the start and costly repetition of work already done can be avoided.

The audio engineers also perform quality control (QC) checks on final archival work (Figure 2) to ensure that the data archived are as accurate and secure as possible. The QC work includes testing the disks destined for deep storage using an Audio Dev CATS system, an industry accepted professional QC hardware/software system. Reports generated by the CATS are retained for later use during periodic verification check of stored DVDs. Verification of stored DVDs is done to determine what, if any, degradation of the media has occurred. QC also requires that preserved digital content be checked at regular intervals for data integrity to ensure that files have not become corrupted and are accurate copies of the original digitised recording.[3] To do so, a checksum strategy is employed, specifically the Message-Digest algorithm 5 (MD5),[2] whereby an MD5 for each sound file is stored within the file metadata header (see below). The embedded MD5 is used to compare and verify the integrity of the stored archive sound files, both on DVD and the SAN.

The end products of the digitisation process ensure long-term integrity of the data by creating multiple redundant copies of each file, stored in multiple locations (Figure 2). First, two DVD-ROMs containing 96 kHz/24-bit files are produced: one master for on-site use and data retrieval and a second safety copy for deep archival, which is stored off-site. Secondly, a high-resolution (96 kHz/24-bit) archival file is also stored in an on-site HDD SAN system utilising the RAID 5 redundancy architecture that is mirrored to an off-site SAN. These actions ensure that the audio data are secure and retained intact long into the future. Equally important is the ability to retrieve the audio files and make sense of their content.

METADATA AND FILE RETRIEVAL

Digitisation of analogue media constituted only one facet of the larger goal of converting the ML into a digitally accessibly research resource. Software tools for searching and finding items in the archive also had to be created, and these tools needed to meet the search and retrieval demands of the contemporary internet user. Among the ML's client base, which ranges from school-aged students and their teachers to artists and entertainers, the largest single body of users of the assets is the scientific researcher. Of paramount important to this type of client, and many other clients as well, are searchable fields for scientific name, geographic information, behavioural data, habitat description and technical specifications of the field recording system. These data are the core elements that comprise metadata records submitted by the contributors to the archive and it is critically important that these metadata be searchable by researchers and other clients. Toward this end, the ML's first computer database was developed in 1978 by then director Dr James L. Gulledge, using the Integrated Scientific Information System (ISIS). In 1997, the ISIS database was migrated to a FoxPro database to meet the need for a database that not only served in-house users but also field recordists who needed to enter data on a laptop while at remote field sites. The year 2007 marked the transition to the library's current Oracle database structure: a relational data model that is more compatible with internet retrieval and also with Dublin Core, Darwin Core and PBS Core. This database also affords straightforward data harvesting by the Global Biodiversity Information Facility (www.gbif.org), VertNet (vertnet.org) and other biological data clearinghouses.

In 2009, ML sought a major revision of archival storage and management capability, not only to address the requirements of the expanding digital audio assets, but also to better serve the needs of a rapidly expanding effort by the library to archive animal behaviour videos. Archival storage was originally in large folders containing up to the OSX directory limit of 100,000 files per folder, but this strategy did not facilitate efficient search. Management of the files was complicated by the fact that primary and proxy assets were stored in the same volumes, less than ideal with regard to security. Accordingly, the archive structure was redesigned, based on each asset's unique catalogue number, and archival storage was expanded to 40 TB. This modification permitted significantly more rapid search and automated (scripted) workflows, watch folders and other semi-automated procedures. For the purpose of improved security, media assets were also separated into categories and separate volumes were created for primary and proxy assets. Following shortly after archival storage redesign, the library deployed Apple Inc's Final Cut Server 1.5 to improve the ability to manage and maintain the archive on a large scale. It is also used by ML staff responsible for serving requests from clients for media files to search, manage and annotate assets as well as project creation. With Apple's Final Cut Server end-of-life announcement, the ML is investigating potential solutions, such as North Plains Telescope and Merlin*One*'s digital asset management system.

Another critical issue was to have the metadata strongly associated with the media specimen and an invaluable step forward was to embed key metadata within the sound file itself. In 2010, ML began using the Broadcast Wave file format (BWF) at 96 kHz/24-bit for its archival audio file format, as this format allows the storage of a limited amount of standardised data, and also the MD5 used for QC, within the header of the wav file.[4] Although this standard data set has limitations in terms of the size of character strings permitted, widespread adoption of this format by audio industry software and hardware companies means that many people have access to the embedded descriptive data. Because the ML database is accessible via the web, it was possible to address limitations on the length of the character string by embedding a web link to the sound file's full data set. Thus, anyone with the ability to read the BWF metadata can access the complete data record for any sound file. One additional advantage of storing the metadata within the file is that it removes the risk of losing the link between metadata and the digital audio, which is potentially important to researchers whose active data set may involve thousands of audio files.

LESSONS LEARNED AND INTO THE FUTURE

Through the efforts to digitise the ML collection, a number of lessons have been learned, which will help ML (and others) prepare for the future. First, it is critical that digital files are stored in a persistent electronic archive with multiple copies at multiple locations and in a format that facilitates future file migration. A clear migration strategy is critical, as changes in digital technologies and file formats are inevitable. Secondly, a digital media archive like the ML, which is a resource dependent on the continuing contributions of others (eg researchers and enthused hobbyists), must embrace new technologies and develop methods to engage recordists of the 21st century, enabling them to contribute their digital media and metadata easily and quickly. With the availability of high-resolution digital recorders and the numbers of individuals using these devices to document wildlife, particularly young researchers, the growth potential for a media archive like the ML is enormous. To meet this need, the library is developing tools that will allow individuals to upload and manage their own recorded assets within the archive. The expectation of these young researchers and volunteers is immediacy in addition to the capability to accommodate the different media formats (audio, video and still images) as well as detailed metadata that are now routinely gathered in the field. Finally, at this same time it is critical to keep another foot firmly planted in the past, namely, by having the equipment and expertise needed to handle (ie digitise to the highest fidelity possibly) analogue media that continue to be contributed by recordists from the latter half of the 20th century. These recordings, often found on the shelves and in the attics of retired researchers, can span decades and are invaluable media specimens; in many cases documenting places, populations and even species that no longer exist (indeed, the ML has several archived recordings of species now extinct). It is critical that a scientific sound archive be prepared to handle this legacy material.

Digitisation of the ML has brought two enormous benefits: long-term electronic preservation of the collection and easy accessibility. Because the public can now explore the assets online at macaulaylibrary.org (over 800,000 unique visitors during the past year), researchers and other clients can find the exact recordings they need for a project and then request high-resolution copies to be delivered via ftp. This has led to exponential growth in requests for material — the number of orders has nearly doubled in each of the past four years. The average size of an order has also increased, with researchers sometimes requesting thousands of audio files for large-scale projects. Use of media assets has also expanded dramatically into other areas as well. For example, thanks to the exposure brought by ML's digitised online presence, artistic uses of audio have included a museum exhibition developed by Paola Pivi, in which visitors can compose music and 'jam' with animal sounds (http://www.grrr.nu/wp/?page_id=192), a series on 'wild sounds' that aired on National Public Radio (http://www.npr.org/series/93588962/wild-soundswild) and even a segment on WNYC's Studio 360 that challenged listeners to remix animal sounds to make music (http://www.studio360.org/2013/mar/01/listener-challenge-remixing-spring/). These examples illustrate nicely the creativity that can spark when artists and scientists alike have access to a digital media collection like the ML.

References

1. International Association of Sound and Audiovisual Archives (2009) 'Guidelines on the Production and Preservation of Digital Audio Objects', 2nd edition, edited by K. Bradley, available at: http://www.iasa-web.org/audio-preservation-tc04 (accessed 23rd October, 2013).
2. Casey, M. and Gordon, B. (2007) 'Sound Directions: Best Practices for Audio Preservation', Indiana University, Bloomington, IN and Harvard University, Cambridge, MA.
3. International Association of Sound and Audiovisual Archives (2005) 'The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy', IASA-TC 03, Version 3, available at: http://www.iasa-web.org/sites/default/files/downloads/publications/TC03_English.pdf (accessed 23rd October, 2013).
4. European Broadcasting Union (2011) 'Specification of the Broadcast Wave Format (BWF) EBU-TECH 3285', Version 2, available at: https://tech.ebu.ch/docs/tech/tech3285.pdf (accessed 23rd October, 2013).

