DOI: 10.2352/issn.2168-3204.2016.1.0.156

# Embedding Metadata in Large-Scale Legacy Digital Audio Collections

*Ryan Edge; Michigan State University Libraries; 366 W. Circle Dr. (13), East Lansing, MI, USA*

## Abstract

*Embedded metadata is widely recognized for its ability to enhance the discovery, usability, and management of digital assets. This paper details the process of establishing and implementing core embedded metadata for a collection of over 23,000 digital audio master files at the United States' largest academic recorded voice repository. Obstacles and solutions discussed in this paper address both the common and exceptional issues in retrospective collection description, including idiosyncratic data sources, systematic metadata transformation, encoding inconsistencies, and file remediation.*

## Background

At the Michigan State University Libraries, audio assets factor heavily into our collections of distinction. The G. Robert Vincent Voice Library (VVL) is a part of the Libraries and is the largest academic recorded voice collection in the United States, containing primary source sound recordings, primarily in speech, interview, lecture, and performance formats. As the majority of the collection originates from the 20th century, and was captured on analog magnetic tape, in-house digital reformatting has been a recognized imperative here for over 15 years. In that time, audio digitization standards have taken root in the field, and the Voice Library has been hard at work. At the time of this writing, there are over 23,000 digital preservation master WAVE files (and counting) in the Voice Library's collection.

Since the beginning, the Libraries' local audio digitization program has striven for archival digital audio standards compliance. Our analog-to-digital reformatting workflow produces uncompressed linear pulse code modulated audio (24-bit, 96 kHz; spoken word: 16-bit, 44.1 kHz) contained in a WAVE file, but like many other organizations, our commitment to embedded metadata was often irregular. As files accumulated rapidly over the years, the notion of retrospective processing for a collection of this size was reasonably dismissed as unduly onerous. Of course the most crucial information was being captured elsewhere, in a processing database which contained tremendous documentation of the recorded content and administrative actions taken with these assets. And therein laid the key to enhancing discovery and management of this vast digital audio collection.

## Embedded Metadata in BWF Audio

The Broadcast Wave Format (abbreviated as BWF or BWAV) specification was unveiled in 1996 by the European Broadcasting Union (EBU) as a standard exchange format for audio files regularly passed through different computer environments and asset systems. For this reason, and others discussed later, BWF has seen strong adoption throughout the audio preservation and broadcast engineering fields as the standard target audio preservation format, recommended by the International Association of Sound and Audiovisual Archives (IASA) [1] and Audio Engineering Society (AES) [2].

The Broadcast Wave Format builds upon the standard WAVE audio format, which itself is a form of the generic RIFF container. Like any RIFF variant, a WAVE file (and BWF by extension) stores data in tagged "chunks." Chunks are compositional sections of the file; each chunk serves a specific purpose and has a specific structure. All valid WAVE files have a RIFF header, which, among other details, specifies the form type code ("WAVE"), file size, and accompanying sub-chunks. The two essential chunks are Data (wave-data) and Format (fmt-ck). The Data chunk contains the raw sample data—an encoded audio signal, typically an uncompressed pulse code modulation (PCM) audio bitstream. The Format chunk specifies the technical properties of the audio data so that it can be correctly interpreted: information about sample rate, number of channels, bit depth, etc. [3].

Although BWF and WAVE are functionally interchangeable and cross-compatible—both bear the same ".wav" file extension (there is no ".bwf" extension)—there are a few differences of particular importance to archives. Besides restricting the contained audio codecs to just PCM and MPEG Layer I/II audio, BWF extends WAVE by an additional data chunk. The BEXT (Broadcast Audio Extension) chunk allows additional metadata to be stored within the structure of the WAVE/BWF file container. BEXT is the de facto standard for embedded archival metadata and the most widely supported across platforms [4]. (Though there are several other metadata carriers [i.e. XML-based chunks: XMP, iXML, aXML], this paper discusses only the LIST-INFO chunk as a suitable complement to BEXT.)

BEXT enables storage of basic administrative and descriptive information in each audio file: the original file name, the collection to which it belongs, known identifiers and aliases, file use designation (e.g. preservation), date of file creation, source format (e.g. audiocassette), equipment and settings used in the analog-to-digital signal chain, in addition to what entity claims archival responsibility. The core BEXT metadata set more or less corresponds to those facts that are forever relevant and unchanging. And the values placed within them are highly restricted: most fields have an ASCII character limit and are strongly recommended to abide the vocabulary and syntax parameters put forth by the EBU and the Federal Agencies Digitization Guidelines Initiative (FADGI).

## Defining Scope & Methods

In order to accelerate this massive legacy digital collection toward BWF compliance, and to do so responsibly, we had to first test our assumptions and tools we intended to use. Media Preservation began by forming a profile of the existing embedded

 

DOI: 10.2352/issn.2168-3204.2016.1.0.156

metadata in the Dark Archive, the storage area for our preservation master collection. Using ExifTool's batch metadata export (-csv) feature, we were able to review all of the Dark Archive's WAVE files and attributes in aggregate. This helped to inform the group about what we stood to gain, as well as the data residue already contained in these files. Test duplicates were then generated for all of the WAVE files and placed in a separate staging directory on the server, where we could safely demonstrate the systematic transformation of preexisting database values and the insertion of these new metadata elements into WAVE/BWF files. Once complete, the project stakeholders could evaluate outcomes and hone the data for final deployment across master files in the Dark Archive.

### *Limits of Embedded Metadata*

In this planning phase we were mindful of retaining ancillary metadata already carried within the file (i.e. recording software and hardware signatures), as a substantial amount of the files were born digital and therefore already held some amount of self-documentation. Of course the more contextual information we have the better, especially if it comes directly from the capture source or signal chain. Appropriate usage of BEXT is inherently restrictive, so therefore, as work progressed, we collected requirements for a secondary embedded metadata carrier with the hope that it might accommodate details that strict, well-formed BEXT simply could not fit.

Embedded metadata is not a substitute for separate XML metadata records, a database, or digital asset management system. Though embedded metadata is adequate for many simple situations, there are considerable limitations to what it can express. Rather it is something like an insurance policy, storing minimal "catastrophic metadata" [5]. Should a file become separated from its external metadata, or a file be altered in some exchange scenario, there is then a key to reestablishing that file's identity.

In addition to this fail-safe component, the Library was interested in embedded provenance and attribution data. This information is often lost as files are decontextualized through regeneration and interchange across unknown future platforms and workflows. The Voice Library also wanted to ensure that key elements were going to be passed down from the masters to their eventual MP3 access derivative files. So in addition to BEXT, we explored supplemental metadata sets to make appropriate use of the abundant data at our disposal.

### *LIST-INFO as a Secondary Metadata Carrier*

Ultimately LIST-INFO was selected as the metadata chunk for our purposes. It is the second best supported metadata set (following BEXT) and is advocated for use in combination with BEXT by FADGI, which provides clear documentation on its utilization. Though LIST-INFO elements are well defined, usage is more flexible than it is for the BEXT set: no character limits, no prescribed syntax (see EBU R98-1999 constraints on BEXT CodingHistory construction), and some interpretive latitude. The Voice Library was also pleased to have several more useful parking spaces for high-priority descriptive data points (e.g. Speakers, Date of original recording), and the capacity to pass down several key elements to equivalent MP3 ID3 tags in eventual derivative access copies, with a tolerable degree of semantic shifting.

### *BWF MetaEdit*

Developed by FADGI with support from AudioVisual Preservation Solutions (AVPreserve), BWF MetaEdit is a free, open source application that enables editing, embedding, and validation of metadata in WAVE/BWF files in accordance with the FADGI guidelines [6], specifically in the core BEXT and INFO chunks. With the set of functionalities MetaEdit provides, any organization with existing collection data can substantially upgrade its archival digital audio assets. Batch import through comma-separated value (CSV) files makes inserting metadata from an external source into a set of audio files simple. The most significant hurdle, of course, is shaping one's data pool to a digestible interchange format for use by MetaEdit.

### *Text Wrangling: A Group Activity*

As with any large-scale retrospective data processing project, this effort required a few distinct skill sets and cross-departmental collaboration. Media Preservation depended heavily on the technical proficiencies of the Digital Library Programmer and the Systems unit; their involvement was critical: enabling access to the Voice Library database, querying the database, and then converting the extracted values into the rich, well-formed BWF metadata that could be deployed with MetaEdit.

## The Data Source

Historically, all Voice Library collection data has been stored in an SQL database initially designed for finding aid development, rights determinations, and content summaries. Over time, it has been augmented to accommodate many more processes and notes. Working with the Voice Library's collection manager and long-time engineer, we were able to delineate quality data points from outmoded or low-confidence data and began reconciling our most trustworthy database elements with FADGI's recommended fields. At this time we also earmarked additional database fragments that we might be able to refine and concatenate in order to construct meaningful BEXT/INFO values.

As the more than 23,000 principal files are products of long-sustained reformatting operations (over 15 years and counting) involving dozens of technicians and created under evolving technologies and practice, the database was known to contain some natural data entry errors and inconsistencies—names, titles, etc. This much was granted. There was also tremendous redundancy due to the design of the database. In most cases, there were multiple database records for a single recording: one for each manifestation, for example, the original open reel recording, all physical duplicates, and the digital preservation master. Among like records, there were often unique details or notes regarding the origin or administrative lifecycle of the asset. We could not just discard this data, so we looked for patterns to harness in subsequent programmatic data processing efforts. Naturally data cleanup played a large role; we merged and reformatted the data to suit our needs, but were also mindful of the rich information ecosystem that we were exploiting, and the implicit relationships that we could jeopardize.

DOI: 10.2352/issn.2168-3204.2016.1.0.156

## Shaping and Rendering Data

Once the provisional metadata elements and remediation measures were plotted, the SQL database was queried to retrieve the necessary records. In this first pass, far-reaching joins and normalization were incorporated to set the stage for subsequent programmatic processing. Inconsistencies and undesired data formats were also remedied at this time (e.g. "Jan 1 2000" to "2000-01-01"). The CSV output was then reviewed for properties that might inform the subsequent transformation processes. In seeing the data in a digest CSV spreadsheet form, new handles on the data emerged with greater clarity.

### *Iterative Transformation*

The most extensive transformations came through the Python script (developed by our Digital Library Programmer), which parses the CSV line by line, coupling associated records based on the unique identifier assigned to the recording and its derivatives. Using a simple heuristic method to determine what sort of entity each record/row represented (e.g. master or derivative copy), the script then routes each element of the record into the appropriate BEXT and LIST-INFO field. Related records are also harvested for unique data, which are added into the primary record, thus producing details on the recording's manifestations and known aliases. Conditional formatting is also applied to bring structure to complicated BEXT fields like Description and CodingHistory.

A number of operations contained within the transformation script were reevaluated as the project team negotiated parameters of the final metadata set. Due to a loose timeframe and lack of a project charter, however, a certain degree of function creep resulted. The iterative reformatting and review processes often meant having to defensively control stakeholders' expectations and mitigate temptations to work around best practices and to sacrifice accuracy for the sake of fitting one or more tokens of information that, although valuable, were not suitable as semi-permanent embedded metadata. The following section highlights this dialogue in greater detail: the allowances, challenges, and constraints of our particular data source—what we were able to construct and what we chose to throw away.

## Enhancements and Exclusions

Owing to the deliberate pace of the project, we were able to take time to crawl seemingly peripheral data fragments, and extrapolate from them content appropriate for our metadata. For instance, in the event of records with the same content ID, all associated format types designated as physical formats were carried into the IMED field, in order of generation, and delimited by semicolons.

Another option we pursued was the automated description of past signal processing (appropriate for the BEXT CodingHistory field). Because the Voice Library kept notes on past reformatting processes and equipment, we were able to programmatically construct a generalized reformatting process history for each recording based on format type and era of transfer (e.g. device X was solely compatible with format Y between dates A and B, therefore signal chain is Z). This function was abandoned, however, due to an inability to confirm this complex data's quality at our collection's large scale.

Throughout the project, the working credo was "first do no harm," often meaning that we would rather see an asset go untagged or under-described than tethered to untrustworthy data. For this reason, the CodingHistory field for most extant WAVE files contains only basic information about the original recording, and little on the transfer processes and technologies used to create the digital copies. However, signal chain and capture specifications are currently being woven into the next iteration of the database, so now all newly generated WAVE files will include this data, formatted as CodingHistory indicates in Table 1.

Although the VVL database supplied us with nearly all necessary information for the BEXT and LIST-INFO sets, there were several INFO elements that we elected to forgo. For instance, names of student workers responsible for reformatting actions were deemed excessive or inappropriate, and thus excluded (from the Engineer [IENG] and Technician [ITCH] fields). Data like this is of course important for a number of administrative reasons but is documented elsewhere. Likewise, copyright statements and usage restrictions are perhaps even more significant, and valuable to end users, but the Copyright (ICOP) element remained unaddressed due to the ever-shifting status of many of the recordings (a boilerplate rights statement to reflect this is currently under review). Subject (ISBJ) and Keyword (IKEY) fields were left unaddressed simply because past usage was deemed reductive or extraneous. And in many cases we had desirable information at hand but not in a consistently structured form that would reasonably allow its programmatic retrieval.

The VVL database contained a few distinct date elements. The LIST-INFO Creation Date (ICRD) tag is intended to store the date of the original recording, and fortunately there was a verified database field from which to extract this date. BEXT features OriginationDate: a field strictly designated for the date of the digital file's creation. This, too, was critical information that the Voice Library counted on, especially for administrative purposes. But unfortunately there was no VVL database field that captured the digitization date. One possible solution was explored: Unix-style file systems record timestamps for three kinds of file manipulation events: date/time last accessed, last modified, and last changed—but no stable date of creation. Though we were able to retrieve any of these file attributes (using the 'stat' Unix system call), they could not be depended upon to represent file origin: anytime a file's inode data is modified, these timestamps change. Because we could not trust that the most recent modification occurred on the date of creation, we could not in good faith embed these date values. Furthermore, editing file header metadata alters the files' data structure, causing the system to overwrite the only record of their digitization vintage. Therefore, as a stopgap measure, before proceeding, the VVL database was amended to include each master files' date of last modification. Though this faulty value was deemed inappropriate for inclusion among our BEXT metadata set, its consideration led us to discover a crucial gap before it was too late.

## Batch Embedding

Once the transformation features were agreed upon, completed, and the intermediary output verified, we segmented the CSV into more manageable batches for BWF MetaEdit import. Little was known about MetaEdit's permitted volume of import,

 

DOI: 10.2352/issn.2168-3204.2016.1.0.156

but it was assumed that anything more than a few thousand records/files might be a challenge. Starting with modest count batches, we eventually were able to ramp up to as many as 10,000 records/files per batch, though the buffer time was tremendously long. Pushing the import limits to this volume is not advised; we were only willing to take the risks because we were at this time working with test duplicates, not our preservation masters.

Before repeating the embedding process with the Dark Archive master files, we reviewed the test duplicates' embedded metadata again using ExifTool's batch metadata export feature. This proved to be very useful as a means of presenting the exact impact of the process in an uncomplicated spreadsheet format. All stakeholders were given the opportunity to review and verify the integrity of the data before finally pulling the trigger.

**Table 1: Select BEXT and LIST-INFO elements as applied at the Vincent Voice Library**

| Element | EBU/FADGI Recommendation | VVL Application / Example |
|---|---|---|
| *Description* (BEXT; 256 char limit) | Collection ID; Source object ID. File use. Title control number. Original filename | Collection VVL Item M5374 bd.12. File use: Preservation master. Original filename: DB28697.wav |
| *Originator* (BEXT; 32 char limit) | Entity responsible for "archiving" the audio content | Vincent Voice Library, MSU |
| *CodingHistory* (BEXT; unrestricted char limit) | Data describing reformatting process. See EBU-TECH R98-1999 [7] for proper syntax.<br>Coding algorithm ("A="); Sampling frequency ("F="); Bit depth ("W="); Mode/Channels ("M="); Text ("T=") | A=ANALOG,M=mono, T=Otari MX-5050; 7.5 ips; open reel tape,<br>A=PCM,F=96000,W=24,M=mono,T=Yamaha 01V96; ADC1,<br>A=PCM,F=96000,W=24,M=mono,T=E-MU 1616; DIO, |
| *IART* (LIST-INFO) | Artist(s): creator or contributor to the original content | Speaker(s): Wharton, Clifton R.; Rustem, William |
| *ICMT* (LIST-INFO) | Comments: provides general statements about the file or subject of the file | Summary: Clifton Wharton speaks in a news conference following student demonstrations on . . . |
| *ICRD* (LIST-INFO) | Creation Date: date of the original recording; ISO 8601 format | 1970-02-20 |
| *ISFT* (LIST-INFO) | Software: package used to create the file | Adobe Audition 4.0.0.1815 |
| *ISRC* (LIST-INFO) | Source: person or entity who supplied the content of the file | Gift of Fred Bruflodt |
| *ISRF* (LIST-INFO) | Source Form: original form of the material that was digitized | Open reel audio |

## Unwelcome Discoveries and Remediation

Ultimately the unexpected outcomes proved to be among the most striking and urgent of the project. In the batch embedding process, 84 files with various dysfunctions were discovered, their file headers being cited as inaccessible by MetaEdit. Fortunately, the application throws an exception for each invalid file, along with a brief message regarding the nature of the error (e.g. "truncated"; "fmt_"; "no RIFF/RF64 header"). Following these clues, we examined each file through playback or characterization tools (MediaInfo, FFprobe, ExifTool) in order to determine viability and identify encoding anomalies. In many cases, each of the tools yielded contradictory results or wildly unusual file attributes (e.g. 16 kHz and 88.2 kHz sampling rates, 4 and 7.1 channel audio). Inconsistency is not uncommon; a system's interpretation of a file's properties is often only as accurate as the file's self-documentation. But this caused us to look further, and what we found was varied and surprising.

### *RIFF Remediation*

Files identified by MetaEdit as having no declarative RIFF chunk could not initially be played back. This is not because the WAVE is missing a header; rather, it has a flawed structure. To rehabilitate these files, Audacity and FFmpeg (in separate processes) were used to extract the audio stream and repackage the data in a "new" WAVE container. In these cases, Media Preservation was able to restore the content as fully functional BWF preservation quality files: WAVE audio capable of playback and headers capable of accepting embedded metadata.

### *Truncated Files*

Unlike the problematic RIFF header files, playback for the reported "truncated" files was unaffected. The audio data itself is intact; we simply cannot edit the header due to an inaccurate chunk size declaration (file size is less than stated in the chunk size declaration, which is typically the result of interrupted duplication or transfer [8]). Assuming that these files had been truncated upon duplication to the staging directory, we referred to the digital masters only to find that they too were affected and played back without error. The precise local cause remains unknown, but it seems likely that these files were compromised at their conception. However, they are entirely functional, just incapable of receiving embedded metadata at the present time. Remediation plans are currently under review.

### *Remainders*

Of the remaining 15 files, most were born-digital recordings compromised by fatal encoding errors, while the final few were unidentifiable outliers, unexplainable mistakes, or contained a codec incapable of being wrapped in a BWF-compliant WAVE file (e.g. MPEG Layer III, AAC).

## Next Steps

Finally the massive pilot was complete, and the data formatting and embedding processes were refined slightly once more before insertion into the true preservation masters across the Dark Archive. In total, 99.8% of the preservation master files were embedded successfully. Further investigation and recaptures are anticipated to rectify the previously discussed excluded assets. The

next steps are already underway: based on the comprehensive review of the database and lessons learned from this project, Media Preservation is now devising revisions to the VVL database in order to simplify information fields significant to Voice Library workflows, tailored specifically for use in BEXT and INFO sets. In the near term we also intend to consolidate technicians' data entry by populating BEXT/INFO fields, the database, and, by extension, external MARCXML records (near-) simultaneously. Whether this is accomplished at the capture stage or post-capture, through an additional ingest script function, has yet to be determined.

## Conclusion

Any archiving organization can accomplish some level of retrospective metadata reconciliation, regardless of scale, to the degree that there is relevant data to apply. While the Voice Library has clearly benefited from a robust legacy database, there are many other options that can range from stopgap measures to strict application of best practices. Ideally, organizations would begin to embed metadata tomorrow, using one of the many common sound editing software; Adobe Audition and Steinberg WaveLab support BEXT and INFO tagging, while Audacity supports access to most INFO tags. If for whatever reason BWF MetaEdit is undesired, there are many other adequate metadata embedding methods supported in consumer-grade digital asset management applications as well as freely available post-recording tag editors.

If one is fortunate enough to have XML records for their digital assets, XSL transformation can be used to output data into any number of interchange formats to either develop a database or begin batch embedding. For loose or uncollected legacy data, there may be potential to glean useful information from assumed miscellanea stored in multiple file locations. Discussion with local system administrators and programmers can define possible courses of action; be clear about your intended outcomes; to many who work in common information processing, the compositional tasks you propose will likely resemble their own work.

By unifying critical collection data, nearly any minimally structured documentation (e.g. database, spreadsheet, XML) can serve as your data's life raft as you develop a long-term strategy. It is the author's hope that this paper reveals the low bar to strengthening value and usefulness of digital media, and encourages others to revisit their digital collections' sleeping giants, regardless of scale.

## Acknowledgements

Special thanks to Nathan Collins, Information Technologist; Devin Higgins, Digital Library Programmer; Rick Peiffer, Engineer, Vincent Voice Library; John Shaw, Head of the Vincent Voice Library.

## References

[1] IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, IASA-TC 04, 2nd ed. (2009). Retrieved Jan. 12, 2016, from www.iasa-web.org/tc04/audio-preservation

[2] Audio Engineering Society, AES Standard for Network and File Transfer of Audio - Audio-File Transfer and Exchange - Part 3: Simple Project Interchange, AES31-3-2008, 3rd ed. (2008). Retrieved Jan. 12, 2016, from http://www.aes.org/tmpFiles/aessc/20160117/aes31-3-2008-r2013-i.pdf

[3] European Broadcast Union, Specification of the Broadcast Wave Format (BWF), EBU Technical Specification 3285, 2nd ed. (2011). Retrieved Sept. 8, 2015, from https://tech.ebu.ch/docs/tech/tech3285.pdf

[4] ARSC Technical Committee, Study of Embedded Metadata Support in Audio Recording Software, (2011). Retrieved Sept. 14, 2015, from http://www.arsc-audio.org/pdf/ARSC_TC_MD_Study.pdf

[5] The Audio Archive. (2014). Audio Metadata Primer, (2014). Retrieved Sept. 21, 2015, from http://www.theaudioarchive.com/TAA_Resources_Metadata.htm

[6] Federal Agencies Digitization Guidelines Initiative, Embedding Metadata in Digital Audio Files, 2nd ed. (2012). Retrieved Sept. 8, 2015, from http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Guideline_20120423.pdf

[7] European Broadcast Union (EBU), Format for the <CodingHistory> field in Broadcast Wave Format files, BWF, EBU Technical Recommendation R98-1999 (1999). Retrieved Jan. 10, 2016, from https://tech.ebu.ch/docs/r/r098.pdf

[8] Federal Agencies Digitization Guidelines Initiative, Embedding Metadata in Broadcast WAVE Files - BWF MetaEdit Help: Explanations of Errors and Warnings. (2010). Retrieved Sept. 17, 2015, from http://www.digitizationguidelines.gov/audio-visual/documents/errors.html

## Author Biography

*Ryan Edge is the Media & Digital Preservation Librarian at Michigan State University Libraries, where he coordinates audiovisual reformatting, born digital content migration, and digital preservation activities. Prior to this, he served as Project Manager for the IMLS-funded Preservation Self-Assessment Program (PSAP) at the University of Illinois Libraries. He received his MS from the University of Illinois Graduate School of Library & Information Science in 2013.*

 