EMBEDDED METADATA IN WAVE FILES: A LOOK INSIDE ISSUES AND TOOLS

Chris Lacinak, AudioVisual Preservation Solutions, USA[19]

What is Embedded Metadata

Metadata is an integral component of digital preservation and an essential part of the digital object. Files without appropriate metadata are not understandable, interpretable, or manageable. Effectively, there is no preservation or meaningful access without metadata.

In the file-based domain metadata can be stored using one of two primary methods. First, it can be stored external to the digital object, such as in a database or XML file. This external record must necessarily be associated to the source object via file pathways, unique identifiers, or other means. Second, metadata can be embedded with the object itself. "Embedded metadata" can be most simply defined as metadata that is stored inside the same file, or container, that also stores the essence to which the metadata refers.

In many ways one can think of embedded metadata as the file-based domain's equivalent of labels, annotations, and written documentation stored inside of material housing, or even as "in-program" annotations such as audio and video slates at the head of a recording. In essence, these are all means of maintaining external or non-inferable intellectual knowledge pertaining to content and carrier, leveraging the technology supported by the carrier.

This article discusses the significance of embedded metadata and some of the challenges that arise once this understanding is realized. While the discussion ranges across media types and file formats, the focus is more specifically on metadata embedded in audio WAVE files.

Why embed

In 2009, the Federal Agencies Digitization Guidelines Initiative's Audio Visual Working Group (FADGI) published a document titled, *Embedding Metadata in Digital Audio Files*[20] that explored the topic of embedded metadata within the context of archives. The following is an excerpt from this document:

> **"Why embed metadata?**
> Embedded metadata can provide information to and support functionality for various persons and systems at a variety of points in the content life cycle. For example, it can help the digitizing unit or organization as it produces and preserves content. It can serve persons or systems who receive content that is disseminated by the digitizing unit or organization. Some metadata elements are especially valuable to internal actors, some to external, and some to both.
>
> Embedded metadata, of course, is rarely an agency's only metadata. In most archiving and preservation programs, workflow and archiving are supported by one or more databases, cataloging systems, finding aids, and the like, each of which contains metadata."

For those already embedding metadata in files, the primary intent is generally to convey only the critical information necessary for intellectual and access control in situations where the

19 AudioVisual Preservation Solutions 350 7th Ave., Suite 1603, New York, NY, 10001, USA. ph: 917-548-8632, fax: 866-264-4275, chris@avpreserve.com, http://www.avpreserve.com, facebook.com/AVPreserve, twitter.com/AVPreserve

20 Federal Agencies Audio-Visual Working Group. 2009. *Embedding Metadata in Digital Audio Files: Introductory Discussion for the Federal Agencies Guideline*. http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Intro_090915.pdf Consulted 20 May 2011.

file has been disassociated from its database, maintaining a minimal approach to the number of fields selected for embedding.

Is "why" even the right question?

To be clear, the FADGI report referenced above explicitly states "The associated guideline pertains to embedded metadata in audio files that result from the reformatting of analog content." The question addressed within this stated constraint is, "Why embed metadata?" In other words, why establish workflows outside of the reformatting process to actively capture and insert metadata for embedding with a file?

The reality is that despite any choices about whether or not to embed metadata in files resulting from digitization activities, embedded metadata is already pervasive in born digital media.

Producers of content rely on embedded metadata for management of media and workflows in the production process. This need has resulted in production and asset management developers integrating embedded metadata mechanisms into their systems. End-users rely on embedded metadata for search and retrieval of audiovisual content on a daily basis within computer operating systems, applications such as iTunes, and on sites such as archive.org where users share audio files.

These requirements for producers and end-users often converge in the distribution and use of file-based materials. As a wide spread example, it would be almost unthinkable for a band to place audio files online without embedding metadata in them to at least identify the artist, song title, rights, etc. Without this information, users could not find and access songs, and distributors like iTunes would have no means of determining what song was sold and, therefore, no means of assigning and disbursing royalties.

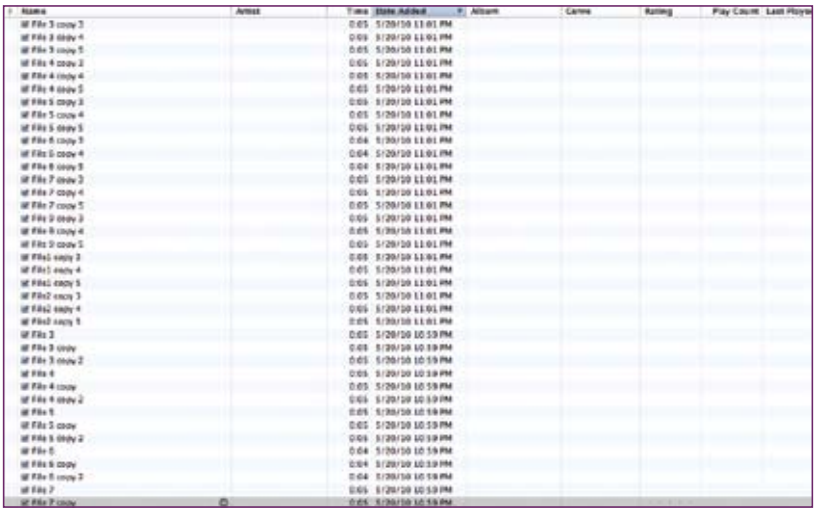

Figure 1: An example of what an iTunes library would look like if files imported had no embedded metadata

Looking outside of the audiovisual realm, the legal system has also taken note of the significance of embedded metadata. In a 2009 Arizona Supreme Court Decision,[21] the Court ruled in favor of a police officer that sued for employment discrimination. A letter drafted by his

21 Arizona Supreme Court. David Lake v. City of Phoenix. Arizona Supreme Court No. CV-09-0036-PR. http://www.supreme.state.az.us/opin/pdf2009/cv090036pr.pdf Consulted 20 May 2011.

superiors regarding his performance was submitted as evidence in court. Through inspection of embedded metadata it was discovered that the letter had been backdated and was a falsification. As part of the Court's decision it stated "if a public entity maintains a public record in an electronic format, then the electronic version, including any embedded metadata, is subject to disclosure under our public records laws."

Figure 2: A look at some embedded metadata within a word document

Since the late-1990s, photographers and imaging professionals have worked together to establish and widely adopt the International Press Telecommunications Council (IPTC) Photo Metadata Standard, a robust standard enabling rich sets of embedded metadata to be leveraged within workflows and applications that handle digital images.

Figure 3 shows a picture published on the White House's Flickr account.[22] Figure 4 shows some of the metadata embedded in the photo extracted using Phil Harvey's ExifTool.[23] Figure 5 demonstrates that, using the Mac OS search utility, this photo can be found on a hard drive by searching for the photographer's name. Embedded metadata enables discovery where file names and text (or lack of text, as in image files) are inadequate for search requirements.

Figure 3: Picture downloaded from the White House's Flickr Account

22 White House Photo Stream. Flickr: http://www.flickr.com/photos/whitehouse/5245486755/sizes/l/in/set-72157625560847260/ Consulted 20 May 2011.

23 ExifTool Website: http://www.sno.phy.queensu.ca/~phil/exiftool/ Consulted 20 May 2011

Figure 4: Portion of information embedded in photo displayed in Figure 3 using Phil Harvey's ExifTool

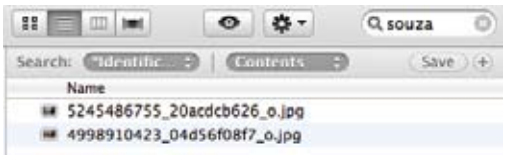

Figure 5: Results of files containing "souza", including the photo shown in Figure 3 and another photo from the White House's Flicker account

Recently, an open source application by the name of *creepy* was developed by Yiannis Kakavas, an academic interested in information security who wants to raise awareness about the information that people are making available about themselves via publishing files to the Internet. Creepy is described[24] by the creator as "an application that allows you to gather geolocation related information about users from social networking platforms and image hosting services. The information is presented in a map inside the application where all the retrieved data is shown accompanied with relevant information (i.e. what was posted from that specific location) to provide context to the presentation." The application harnesses the GPS data embedded into photos and recordings created on personal devices and by services such as Twitter and Foursquare. When made so readily accessible and viewed in the aggregate many people find the level of information disclosed by use of creepy to be disturbing (and thus an aptly named application). This project also recalls the April 2011 discovery that Apple iPhones have been tracking and storing up to a year's worth of GPS location data.[25]

While privacy concerns are not the intent of this paper, creepy is worth noting as a project centered around embedded metadata that has received much attention as of late. That and the iPhone brouhaha also underscore the great need for understanding and being aware of the existence and potential uses (or misuses) of automatically generated embedded metadata.

For those charged with managing collections that include born digital media, awareness of and support for embedded metadata is imperative. Embedded metadata is an aid to archives and conservators in the digital age for verifying authenticity, integrity and provenance of files. It also enables increased accessibility and administrative capabilities. Practically speaking, to ignore embedded metadata is to effectively de-catalog an asset. Metadata capture has long been a resource intensive and challenging proposition for archives. The unprecedented rate of growth

24 Creepy. 2011. Creepy Website: http://ilektrojohn.github.com/creepy/ Consulted 20 May 2011

25 Arthur, Charles. 20 April 2011. 'iPhone keeps record of everywhere you go.' guardian.co.uk. http://www.guardian.co.uk/technology/2011/apr/20/iphone-tracking-prompts-privacy-fears Consulted 20 May 2011.

in audiovisual content production that has taken place in the past decade creates challenges on a completely different order of magnitude. However, with the increased utilization of embedded metadata by systems that automatically populate files with technical, structural, descriptive, and administrative metadata there lies new potential for its efficient capture and use, greatly reducing challenges related to both quantity and quality of metadata.

Tool development

After FADGI established that embedding a core set of metadata in WAVE files was of interest, they needed to decide where to embed the metadata within the file. A study was commissioned and performed by AudioVisual Preservation Solutions exploring and comparing the options for embedding the FADGI proposed metadata, resulting in a report and recommendations.[26] The table seen in Figure 6 below is an excerpt from this report and shows an overview of the comparative analysis performed on the embedded metadata standards reviewed.

Chunk	Size	Definition	Adoption	Authority	Extensibility	Storage
bext	Highly Limited	Highly Limited	High	EBU[xii]	None	Must be before data chunk, at the head of the file.
LIST INFO	Flexible	Limited	Somewhat High	Microsoft[xiii]	Unclear[xiv]	Must be before data chunk, at the head of the file.
iXML	Highly Flexible	Limited, but Extensible	Moderate[xv]	Collection of corporations, website maintained by Gallery	High, may be expanded as needed (registration encouraged)	May appear in any order with the other chunks of the RIFF structure
XMP	Highly Flexible	Somewhat limited, but extensible	In Development within Adobe Products.	Adobe	High, may be expanded as needed (best practices provided)	May appear in any order with the other chunks of the RIFF structure
aXML	Highly Flexible	Highly Flexible	Not commercially available. Apparent internal custom uses within organizations.	EBU[xvi]	Very High	May appear in any order with the other chunks of the RIFF structure, requires the file to meet BWF specifications

Figure 6: Table from page 17 of the referenced report. It should be noted that iXML is currently going through the standardization process within the Audio Engineering Society.

Using this report, FADGI decided to utilize a combination of bext and LIST INFO chunks to meet their needs. Their decisions regarding what metadata to embed and where to embed it are detailed in a document published on the FADGI website.[27]

As a follow-up activity, AudioVisual Preservation Solutions was commissioned by FADGI to develop a free, open source tool that would allow embedding, editing and exporting of metadata within WAVE files. This tool was named BWF MetaEdit and can be found for download at http://sourceforge.net/projects/bwfmetaedit/

Features of BWF MetaEdit include:

26 AudioVisual Preservation Solutions. 2009. Task 5.4: Assess Options for Embedding Metadata in WAVE Files and Plan the Audio Metadata File Header Tool Development Project: Assessment Report and Initial Recommendations.
Federal Agencies Audio Visual Digitization Working Group, Washington D.C. http://www.digitizationguidelines.gov/audio-visual/documents/AVPS_Audio_Metadata_Overview_090612.pdf Consulted 20 May 2011.

27 Federal Agencies Audio-Visual Working Group. 2009. Embedding Metadata in Digital Audio Files: Guideline for Federal Agency Use of Broadcast WAVE Files. http://www.digitizationguidelines.gov/audio-visual/documents/Embed_Guideline_090915r.pdf Consulted 20 May 2011.

- Import, edit, embed, and export specified metadata elements in WAVE audio files
- Batch and individual operation
- Export technical metadata from Format Chunks and minimal metadata from bext and INFO chunks as comma-separated values and/or XML, across a set of files or from individual files
- Evaluate, verify and embed MD5 checksums, as applied to the WAVE file's data chunk (audio bitstream only)
- Enforce specifications developed by the Federal Agencies Audio-Visual Working Group,[28] as well as specifications from the European Broadcasting Union (EBU), Microsoft, and IBM
- Report certain errors in the construction of WAVE files
- Interface through command line and GUI, for Windows/PC, Macintosh OS, Linux. (Full list of OS/interface options reviewable at SourceForge)

Aside from offering capabilities beyond other available tools, BWF MetaEdit is a metadata-centric tool designed to change the landscape of how organizations work with embedded metadata in WAVE files. It is a lightweight, cross-platform tool that can be deployed throughout an organization and used by all stakeholders in the lifecycle of an audio object. Capabilities that were once restricted to specialized audio-centric software usually found only in the audio studio are now made available to everyone, greatly optimizing expertise, increasing efficiency and improving quality assurance of embedded metadata in WAVE files.

Putting embedded metadata to the test

With the new availability of a tool like BWF MetaEdit, AudioVisual Preservation Solutions spearheaded a study in 2010 on behalf of the Association for Recorded Sound Collections Technical Committee (ARSC TC), evaluating the support for embedded metadata within and across a variety of audio recording software applications. The study sought to answer two primary questions:

1. How well does embedded metadata persist, and is its integrity maintained, within any given file as it is handled by various applications over time?
2. How well is embedded metadata handled during the process of creating a derivative?

Three tests, described below, were designed for this study, each one consisting of a test method and reference files.

1. Interoperability and Semantic Shifts: This evaluation uses a reference WAVE file containing extensive embedded metadata within the bext, LIST-INFO, axml, XMP and iXML chunks. The reference WAVE file is opened in a number of software applications to evaluate which fields are displayed in the application's interface. In addition to this, any semantic shifts are documented. Semantic shifts are defined, here, as occurring when an application displays a field's value using a field name that differs from the intent of the original field. As an example: if an application presented the value in the LIST-INFO chunk field labeled "itch" (the term for the technician who digitized the audio) as a field now labeled "artist," this would be considered a semantic shift. It should be emphasized that this test analyzes metadata display only.

2. Persistence and Integrity Through Editing Operations: A four-part test, this evaluation considers how various applications handle embedded metadata when basic metadata and audio editing operations are performed and the file is saved. The first two subtests analyze the results of editing and adding embedded metadata and saving the file.

28 Federal Agencies Audio-Visual Working Group. 2009. Embedding Metadata in Digital Audio Files: Guideline for Federal Agency Use of Broadcast WAVE Files

The third sub-test analyzes the results of performing an audio edit and saving the file. The final sub-test analyzes the results of simply performing the "save as" function. The primary focus for evaluation in all tests is identifying whether existing metadata persists unaltered.

a. Impact of Editing Existing Chunks: This test analyzes the results of editing embedded metadata and saving the file.
b. Impact on Existing Chunks when Creating New Chunks: The test analyzes the impact to existing chunks in a file when new chunks are added and the file is saved.
c. Impact on Metadata of Audio-Only Editing: This test analyzes the results of performing an audio edit and saving the file.
d. Impact on Metadata of 'Save As' Function: This test analyzes the results of performing the "save as" function.

3. Persistence and Integrity Through Derivative Creation: This evaluation tests how various applications handle embedded metadata, when creating a derivative file from a WAVE file. Target derivative file formats tested in the ARSC study included: MP3, FLAC, and WAVE.

Results from this study demonstrated that few of the standard metadata chunks are supported in their entirety by any software application. Rather, applications tend to display and provide access to selected fields of their choosing from each chunk standard. In general, the Broadcast Wave Format bext chunk is the most widely supported, followed by selected fields within the LIST-INFO chunk. Least supported were the XML-based chunks: there was some support for selected fields within iXML, but no support for axml, and support for XMP only by the its creator, Adobe.

Most troubling were the findings associated with application metadata management. According to our findings, it seems to be the rule rather than the exception for applications to automatically *erase* chunks and fields that they do not support after common user actions such as metadata or audio editing are performed. Embedded metadata does not persist nor is its integrity maintained consistently when the same file is used across the audio software applications studied.

These issues have major implications for the use of embedded metadata over time and across workflows. The findings of this study raise serious concerns, particularly for the archiving and preservation communities who rely on embedded metadata for interpretation and management of digital files representing preserved content into the future.

Next Steps

There are two things that readers should NOT take away from this article:

1. The situation is so bad that I shouldn't worry about embedded metadata in WAVE files.
2. I'll wait until it is figured out to do anything.

As mentioned earlier, caretakers of collections ignore embedded metadata at the risk of sacrificing knowledge and utility for themselves and their users. It is important to recognize the current challenging environment so that you are able to navigate it accordingly. While the ARSC TC findings are unsettling on their face, the field of photography offers a hopeful and valuable precedent for successful collaboration between software developers, manufacturers and users in support of embedded metadata generation and persistence. In particular, the Society of American Archivist's (SAA) Photo Metadata Project[29] is an

29 SAA Photo Metadata Project: http://www.photometadata.org/ Consulted 20 May 2011

impressive initiative and serves as an excellent model to follow. It is the author's opinion that the findings of the ARSC TC study are the result of a lack of communication between these stakeholders in the audio industry. With increased awareness and communication,ac the most troubling of these issues, such as erasure of unsupported metadata, are relatively easy ones to solve. At the time of this writing, the ARSC TC Metadata Study Report is approaching publication via the ARSC website.[30] This study helps quantify the issues and will be an effective aid in communication between the various stakeholders regarding embedded metadata support, management and behaviors within applications handling audio files. For readers that are able to contribute to, and participate in, stakeholder discussions, we ask that you please contact this author to discuss the issues further.

More immediately, you can use BWF MetaEdit and the resources produced from the ARSC TC metadata study to put your new awareness of the issues to work by incorporating new quality control procedures and routines into your audio file workflows. Audio "metadata paths" should be considered and tested in the same manner that audio signal paths are. Aside from utilizing the accessibility of BWF MetaEdit to do more rigorous quality assurance, routine integrity testing is also recommended. The ARSC TC study produced three test methods and associated sets of reference files which, at the time of this writing will be available for download from the ARSC website shortly. In combination with BWF MetaEdit, organizations can use the reference files to test their metadata path when configuring systems and as part of routine maintenance and testing.

The significance and pervasiveness of embedded metadata as an integral part of the audio object demands the attention of those who are charged with managing and caring for file-based audio collections. Over the past decade in the transition to the digital domain the archival community often questioned the implications of this shift from physical media to files. The evolution of our collective awareness and toolset for working with embedded metadata, and grappling with how it fits into the information architecture, serves as a poignant example of the ways in which our understanding and practices must change in order to meet preservation principles and provide meaningful long-term access to collections.

30 ARSC Website: http://www.arsc-audio.org/ Consulted 20 May 2011