

# Federal Agencies AV Working Group

Notes from the meeting held at the National Archives (Archives 1, main building downtown), March 8, 2010.

*Abstract: Eight agencies represented. Discussion of a project to document the specifications for a video preservation reformatting target format based in MXF; about work on an MXF Application Specification planned to proceed under the auspices of the AMWA, with a user group meeting planned for November. Discussion of the status of the BWF MetaEdit tool to embed metadata in WAVE audio files; about the pilot projects to test BWF MetaEdit; and plans to make the application available on SourceForge.*

## Video Format Documentation Project

### Introductory

Carl Fleischhauer (Library of Congress) chaired the meeting, which led off with a discussion of how the Working Group would explore one of the potential digital target formats for the reformatting of video (and indeed, possible every type of moving image content). Some highlights:

- The target format to be explored will feature a *wrapper* (MXF<sup>1</sup>), *picture encodings* (JPEG 2000<sup>2</sup>, uncompressed video, possibly others), other essence elements (e.g., sound), and (to the degree possible) *metadata*. The target format (or format "family") employs various open standards and is not in and of itself proprietary.
- The Working Group's exploration builds on the formatting approach employed in the three federal agencies (NARA, Library of Congress, and Smithsonian Institution) that use the SAMMA device sold by Front Porch Digital.<sup>3</sup> The device is a proprietary product but its formatted digital output is not.
- The Working Group has not evaluated the MXF-JPEG 2000 format in detail and has no recommendation at this time. The approach is sufficiently promising, however, to warrant this specification-drafting project.
- The desired specification can be seen as a set of constraints—the limiting of options—against the broad spectrum of possibilities offered by the broad and option-laden standards MXF and JPEG 2000. The outcome can be compared to the establishment of *profiles and levels* frequently used to manage the use of MPEG-2-encoded video or the various *profiles* established to clarify rule sets for JPEG 2000 applications.
- With MXF, it is customary to refer to the specified constraints as an Application Specification (AS). Each AS tends to be associated with a group of users who have similar use cases.

---

<sup>1</sup> Material eXchange Format, standardized by the Society of Motion Picture and Television Engineers as M377 (and other related specification; see <http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml>)

<sup>2</sup> JPEG 2000 is a standard of the International Standards Organization (ISO) and International Electrotechnical Commission (IEC), developed in collaboration with the International Telecommunication Union, Telecommunication Standardization Sector (ITU-T). See

<http://www.digitalpreservation.gov/formats/fdd/fdd000138.shtml>

<sup>3</sup> <http://www.fpdigital.com/Products/Migration/Default.aspx?mrsc=MigOverview>

- Although MXF ASes may ultimately move to SMPTE for further standardization, it is customary to incubate them under the auspices of the Advanced Media Workflow Association (AMWA; <http://www.aafassociation.org/>). The Library of Congress joined AMWA on behalf of the Federal Agencies Working Group in order to advance this effort to document suitable target formats for moving image reformatting.
- The conventions that guide AS form and structure permit of considerable variation. For the preservation-reformatting use case, the Working Group wants something extensible. We may start with the requirements for reformatting standard definition video, but in the future we must be able to extend our approach to high definition and film scanning at various levels of resolution. One representative of the Library of Congress Packard Campus/Culpeper group explained that that their facility planned to use the MXF/JPEG 2000 approach not only for standard and high definition television but also as a target format when scanning film.
- On behalf of the Working Group, the Library plans to engage an expert in MXF matters to help draft one or more ASes to meet needs of preservation-minded organizations like the federal agencies.

Some give and take followed this introductory discussion; the following bullets present a few highlights:

- One attendee mentioned recent meetings with Hollywood and West Coast technical folks in which they continue to define a mezzanine video format, a high quality format that is a bit less high-end than the master archival copy in a digital vault, but still good enough to serve as the source for lower-resolution distribution copies. The mezzanine format seems to be often spoken of as having a data rate on the order of 200 Mbps.
- Someone highlighted the fact that JPEG 2000 was resolution independent; this makes it a good fit for the varying picture sizes in moving image collections.
- One user of the SAMMA encoder reported that it "digitized what was there," i.e., "not as if all the picture data was present." For example, when compressed MPEG files were the source, the resulting MXF-wrapped, lossless-JPEG-2000-re-encoded files "were about the same size" as the original MPEGs, and not the larger size they would be if the input was uncompressed picture data. This was viewed as a good thing, no significant increase in file sizes.

Fleischhauer introduced the idea of convening a meeting of an ad hoc moving image preservation reformatting user group, partly to cater to the AMWA's orientation toward working with groups of users and with potential vendors. Although the Federal Agencies Working Group would be the formal core, a broader user group from many archives around the world would provide more insight and help with adoption of practices over time. AMIA and IASA are having a joint meeting Philadelphia during the first week in November; this should provide a good venue for a user group meeting.<sup>4</sup>

---

<sup>4</sup> Subsequent to the March 8 Working Group meeting, Fleischhauer reported that a time and date for this meeting had been set: from 3 to 6 pm on Monday, November 1, 2010. Exact location to be determined but probably in the main conference hotel.

## **Status Report: BWF MetaEdit Tool to Embed Metadata in Audio Files**

### **Introductory**

Dave Rice of AudioVisual Preservation Solutions (consultants to the Working Group) gave a talk/demo of the BWF MetaEdit tool in its current state; some minor tweaking is planned before it is released to a broader audience. Here are a few highlights from Rice's talk, the surrounding discussion, and follow-up communications:

- The operation of the application relates to a conformance point document, formed as either a comma-separated-value document (CSV) or in XML (see next bullet). There is no direct support for Excel's formats, although Microsoft Excel can "save as" CSV. In a batch mode, a user can read their pre-existing metadata into the conformance point document or the application can use the document to fill in the metadata in the file. The metadata is applied to the bext and INFO list chunks.
- Regarding the XML expression of BWF MetaEdit values, there is an XML schema definition named `conformance_point_document.xsd`. It can be used to validate a conformance point document made by an external application for use with BWF MetaEdit.
- BWF MetaEdit is also capable of filling in the EBU-established aXML and iXML chunks, but does not have the ability to validate this data. The aXML and iXML chunks are beyond the scope of the Working Group's current recommendations although this is a topic that may be taken up in the future.
- BWF MetaEdit offers what is called the "trace view," which lists the arrangement of all chunks in the file and displays all chunks.
- BWF MetaEdit looks for chunks to be of even byte length; this is required for well-formedness by the WAVE specification. For example, you might have a monaural data chunk that is odd; stereo (interleaved bytes for left and right channels) ought always to be even. Odd length chunks are supposed to be padded out; BWF MetaEdit looks for this, checks for padding.
- If a file header says that the file is supposed to be so many bytes (samples?) in size, but the file is the wrong size in relation to that, BWFMetaEdit reports this discrepancy.

### **The BWF MetaEdit "Rule Toolbar"**

In addition to the elements covered in the preceding bullets, two elements received more discussion. One of these concerned the "rule toolbar" in the GUI interface, where you can choose a rule set to apply (e.g., from the Working Group recommendation, from the EBU specification). BWF MetaEdit prevents the entry of invalid data. Here's an example of what you might encounter. The EBU rules for the expression of dates require that you know the day, not just the month and year. Therefore, if you don't know the day, you ought to select a different rule set.

Regarding the rule toolbar, there was some back and forth about having BWF MetaEdit offer users the capability to shape their own rule sets for a particular job. For example, the desired characteristics of newly delivered files would be compared to a profile set by the user, with failures to conform being reported. One attendee noted that batch checking against profiles would very useful in a Q&A process. After the meeting, Rice said that one feasible approach

might be to have a system that allows the user to set up a regex code (regular expression) for each field. In any case, this topic was deferred to the future.

## **BWF MetaEdit Creation of MD5 Checksums**

The second element that received some discussion (including a bit in a post-meeting conversation) concerned the creation of an MD5 checksum or hash value. Rice pointed out that if your MD5 value is for the whole file, when you change the metadata, the checksum/hash is no longer valid. In the current version of BWF MetaEdit, an MD5 value is generated for the sound recording itself, i.e., the bytes in the "data chunk." This does not change as metadata changes and therefore you could check for the integrity of the sound content using the "original" checksum, even though there may have been other changes made to other parts of the file.

Regarding the data-chunk MD5 value, there are at least two issues regarding standardization or practice:

- Should the checksum document the stored data like a traditional md5 (but on only a portion of the file) or the decoded audio (like the flac fingerprint)? FLAC is a format for the lossless compression of sound; see this discussion of the flac fingerprint as compared to MD5 checksums as applied to whole files:  
<http://wiki.etree.org/index.php?page=FlacFingerprint>
- Where should checksum/hash value be placed? There is no generally accepted place to put this. As a placeholder, AudioVisual Preservation Solutions created a chunk-of-its-own for this data. Does the Working Group have a preference for where to put the hash value? Are there any extant relevant specifications for data placement that ought to be considered?

## **Reports from the BWF MetaEdit Pilot Projects**

A few highlights from three projects:

- NARA
  - Setup and testing in September and October, continued testing thereafter
  - Not yet in the production line, still just trying it out; so far, "we love it."
  - The GUI seemed easy to use, very good for non-technical staff.
  - Sidebar on bext-chunk versioning. Here's a little history: the first EBU specification for bext was for version "0" and then, when the capability of inserting a UMID (identifier) was added, the EBU published the spec for version "1." The NARA digital audio workstations, of recent vintage, write a version "1" bext chunk when a file is created. But NARA does not assign a UMID. When BWF MetaEdit edits this metadata and re-writes the bext chunk, since there is no UMID, it writes it as version "0." No one could say why if change from "used to be 1" to "now it is 0" would present any problems.
- American Folklife Center, Library of Congress
  - Looked at batches of files, some with pre-existing "legacy" metadata, some without. Where there was data, they looked at what ought to change.
  - The tool revealed to them one instance of a damaged file, it had been truncated and the "too little data" report emerged. This was very helpful.
  - Having the tool led them to consider where it might fit in various workflows:
  - Work done internally; work done by offsite vendors; legacy file remediation

- The idea of setting up local rules (discussed earlier) is appealing
- Smithsonian Institution Libraries
  - Assisting the Hirschhorn Museum with the transfer of some interviews with artists.
  - BWF MetaEdit used to check the metadata from a service provider.

### **Plans to Place BWF MetaEdit on Source Forge**

Fleischhauer reported that he had met with the Library of Congress Office of the General Counsel to sort out licensing issues pertaining to BWF MetaEdit and SourceForge. All factors look promising and he hopes that the software can find its way to Source Forge by May.