AS-07 Dilemma-response Essay for listservs

Carl Fleischhauer, June 3, 2013

TABLE OF CONTENTS

What is this document?

This is an essay-form response to a productive and provocative set of comments we received during April and May 2013 after we sent queries to the IASA TC-06 list (for the IASA video guidelines project) and to the list for a special AS-07 interest group established at the joint IASA-AMIA meeting in Philadelphia in 2010. We also sent our queries to the member's technical forum at Advance Media Workflow Association, where AS-07 is being developed.

Appendix A provides the text of our query and appendix B summarizes the responses we received.

Firm responses to the *constraint* question we posed

The main question in our April 15 mailing concerned the embedding in AS-07 files of "stray" *Associated Materials*, like images of the tape box with scribbled notes, a document found in the box, or *Supplementary Metadata*, like the logging *process* data output by the SAMMA video reformatting system. We asked about the level of precision that the AS-07 specification should employ when defining the types of files (containing Associated Materials and Supplementary Metadata) that could be embedded. Should this be carefully constrained or left relatively open?

We asked for community input because our own planning team was divided on the question of constraints and to help us sharpen our sense of end-user and system-manufacturer requirements. The summary of the replies in appendix B shows that the voting ran in favor of constraints. Two precise comments came from individuals representing the commercial side, i.e., software developers and system manufacturers:

- MXF AS-07 is a constraint with intent. The constraint is the message. SMPTE AXF has no corresponding constraints; AXF contains the provenance and file system within the AXF format . . . . By contrast, AS-07 is intended to inform media players and media engines how to process the file, so I must align myself with the "specifiers" who recognize the limits of information that can be passed to those automated processors.
- I will vote with the specifiers . . . What we need and want is compatibility, and I do believe that is best served by strong constraint.

The implied question about *packaging*

We framed our question about constraints *assuming* a potential use for AS-07 files as a container for carefully selected materials that are not the usual payload for a "video file." Our view had been that, although strictly optional, we would design a preservation-oriented MXF file that would allow for the embedding of Associated Materials and Supplementary Metadata, to offer an option for encapsulating a range of closely related content.[1] When you put the AS-07 file in your cyber-vault for long-term management, the primary essences and selected related items would be bound together.

This concept brings us to the general topic of *packaging*, to use the terminology of the Open Archival Information System reference model (ISO 14721:2003), especially Submission Information Packages (SIPs) or Archival Information Packages (AIPs). When we posed our question to the lists, we noted that our AS-07 *lower-case-P* packaging option might not be useful for all organizations. With long-term preservation in mind, we are designing MXF AS-07 to complement important *capital-P* packaging formats like the soon-to-be-published AXF standard from SMPTE, capable of encapsulating any types of files with added elements that provide real support for the management of large content packages in storage media. And we noted--as did many commentators--that other high-level packaging schemes are in use in memory institutions around the world, e.g., METS, BagIt, and basic filesystems. Some implementations of these schemes add encapsulation by using zip or tar.

Vigorous responses to the implied question about *packaging*

The greatest part of the discussion on the lists was *not* specifically about the level of constraint to be applied to embedded Associated Materials and Supplementary Metadata. Instead, the writers commented on the implied topic: is it a good or wise practice to employ a

[1] How would an MXF file carry Associated Materials and Supplementary Metadata? SMPTE's family of MXF standards includes ST 410:2008 (*MXF Generic Stream Partition*), a file element intended (at least at first) to carry text (XML, Unicode, plain ASCII). Its development was motivated in part to accommodate Timed Text, an XML format initially specified by the W3C and further standardized in the SMPTE ST 2052 family, and in SMPTE RP 2057:2011. ST 410 not only describes one SMPTE-approved location for Timed Text (you can alternately put it in the Essence Container) but that standard also states that Generic Stream Partitions can also carry elements that are "unevenly distributed along the timeline or large amounts of metadata that cannot suitably be stored as Header Metadata." Even if you put Timed Text in the Essence Container, associated bit-mapped graphics (e.g., a logo) will be placed in Generic Stream Partitions. Thus Generic Stream Partitions provide good locations for our AS-07 Associated Materials and Supplementary Metadata.

preservation-oriented video file as a container for additional materials, i.e., to use the file to contain things like images of the notes scribbled on the old videotape box?

The voting was generally negative, as indicated in the following bullets. In a bit of face-saving rationalization, we felt that these advisories were more about what a wise archivist should *do* with the file rather than being about *writing a specification* that may include such an optional capability.[2]

- I would embed the MXF in our METS schema, using the relations section to take care of Associated Materials and/or Supplementary Metadata.
- The specification should allow only enough embedded content to link the essence to external assets reliably or to identify the essence in case of systemic failure.
- The success of AS-07 (which is by no means assured) rests on being unambiguously written, uniformly implemented, and widely adopted. Multiplying the demands made on the specification and therefore its complexity seems like a puzzling way to go. Other standardised ways exist to relate associated materials and supplementary metadata, and many bodies will already be constrained by extant institutional preferences in this regard anyway.
- I wouldn't recommend to put anything into the archive file which eventually has to be altered later on. [Supplementary] descriptive metadata, for example, may over time need to be corrected or extra information has to be added.
- Currently I'm thinking that it's best to keep the AS-07 format as simple as possible while preserving the full details of the essence (video, audio, timecode). Having all of these essence-related elements inside the MXF file is useful because MXF is able to handle the specialist details of these elements and provide the internal 'file system' needed to relate them in the time domain. [In contrast,] all other items such as videotape box images, programme scripts, etc., are neither 'specialist' A/V media items nor 'time dependent' and so can be handled by a standard IT 'file system', packaged 'loosely' (e.g. BagIt, METS, etc) or 'tightly' (e.g. TAR) or both, as required.

One writer who was at best lukewarm about embedding Associated Materials and Supplementary Metadata nevertheless sketched a use case in which a well-wrapped file package seemed desirable (boldface emphasis mine):

[2] For comparison, in 2013, several digital preservation specialists are participating in a similar discussion in the case of PDF/A, the "archiving" form of PDF that now includes versions *-1*, *-2*, and *-3*, all governed by ISO standards (ISO 19005-1:2005, ISO 19005-2:2011, and ISO 19005-3:2012). The new PDF/A-3 specification permits the embedding of non-PDF files, covering use cases like "provides the malleable source document for current or future as well as the PDF rendering to support long term retention" or "provides the invoice as actionable data as well as a formatted version for printing and retention," and many others. The potential use of PDF/A-3 in document streams destined for memory institutions has caused some consternation. Some archivists reported encountering organizations with PDF document management systems that also possess odds and ends of video content. Representatives of these organizations were reported to have said something like this: "PDF/A-3 is terrific. We had been wondering how in the world we would package and manage the video clips that are among our official records and now we see that we can embed and bundle video with all the rest of our PDFs." It is this type of comment that has produced anxiety among preservation archivists.

- . . . put yourself in the position of a collection manager in a developing country, or underfunded collection anywhere. Imagine you have, by dint of great persistence managed to raise enough funds from an international grants agency to pay a specialist company to preserve your very valuable (to you) collection of video shot in the 1970s on U-matic and given to your Library/archive. **To deal with the language complexities and the multiple paper files that are the only documentation, the contract supplier offers to scan all boxes, card index and notebooks and relate them to the individual files.** The data base and computer system in your Library is functional and managed by a single individual, but the database is not the most complex or developed and cannot manage the complex relations and multiple levels demanded by a comprehensive collection. **What you need (or think you need), is a single object to manage with all the associated data so that when my collection comes back from the contractor you can manage it without having to deal with building a new system and embedding new data** . . . . You also want it to be compatible and standard as you are going to ask a neighbouring organisation/region/country to store copies of the material to mitigate against disastrous loss.

  Other commentators noted technical issues that may arise with large video files:

- . . . if an archive is preserving 6-hour VHS they'll run into media-spanning issues which are also treated by AXF, not by MXF at all. . . . do we want AS-07 to borrow from AXF [in this regard, probably not a good idea]?

AS-07 team responses to constraints, packaging, and complexity

In our team discussion following the listserv exchange, we reconsidered the matter of constraints, packaging, and complexity. Our internal discussion highlighted some of the following ideas:

- Video content--especially reproductions of historical broadcast tapes--with picture data, multiple sound tracks, timecodes, captions, and more, is inevitably and irredeemably complex. There is no way to produce a "simple" digital-file video recording. We were reminded of the complexity of PDF files, now governed by ISO standards, and widely supported by a range of commercial and open-source tools. Digital archivists view them as worrisome at some level, but accept them as a fact of life in the digital age. (We did not even *mention* geospatial data formats, also a fact of digital life and another worry for archivists.)
- Will constraining the file-types for embedded Associated Materials and Supplementary Metadata make the work of file-making and file-reading easier? One team member argued that the opposite was true, saying that constraints would add to the burdens of the encoding-file-making system: it would have to identify and filter what it is to be embedded. At the moment, we plan to continue drafting in this mood, i.e., having no or minimal constraints but also eschewing any requirement on encoders or decoders beyond, "embed pre-existing entities, give the entities back, and do no harm to the entities."
- In any case, and especially if Associated Materials and Supplementary Metadata are embedded, the specification must require the reliable labeling of individual components, and include metadata to link them into parent and child components. This is something that MXF does rather well.

- Encoder and decoder requirements. We use the term *encoder* broadly to name the systems that encode the main essences and assemble the files, and *decoder* to name the systems that open files and read the main essence elements. Compared to the encoding of an AS-07 file's main essences, we plan to ask for a far lower level of performance regarding Associated Materials and Supplementary Metadata. We assume that the images of the old videotape box or the XML metadata *preexist* as the separate outputs of other systems, e.g., a flatbed scanner or an organization's cataloging software. All the AS-07 encoder has to do is let the operator tag and embed the pre-existing items and all the decoder has to do is not harm what is there, display the creator-inscribed tags, and offer up the items' bits for the taking, to be read in another application.
- We acknowledge that the do-no-harm requirement for encoders and decoders has not always played out perfectly. One team member said, "That's supposed to happen with WAVE files, where the rule for players is 'if you don't understand a chunk, just ignore it but don't delete or corrupt it.' But some software fails to follow the rule." Another member replied, "That is just bad code writing, the community needs to insist on good code."

Thus we ended up differing with the trend of comments we received. Regarding the matter of formatting constraints, we continue to favor a laissez faire position on the Associated Materials and Supplementary Metadata to be embedded, coupled with the approach in the bullet above: encoders need not create these items, just embed the bits, and decoders need not read them, just give them back.

Regarding embedding in the first place, we continue to feel that the AS-07 *specification* should allow for the optional embedding of Associated Materials and Supplementary Metadata. One team member said, "We acknowledge that there are use-cases for 'never embed' and for 'always embed,' and organizations may establish a policy of using one or the other or both." We do see that an effective case can be made against the *practice* of embedding "extras," and for encouraging archives to combine the use of AS-07 (and other file formats) with packaging structures like AXF, METS, and BagIt. But we feel that this encouragement ought to be offered in a separate guideline advisory, and not be used to limit the AS-07 specification-qua-specification.

Other topics

The commentary on the lists covered a few topics that we did not feel were central to this discussion, including comments pertaining to support and methods for file integrity. One writer noted that an additional argument for "pure" instances of preservation files (i.e., minimalist files) is that simpler files would make it easier to carry out file-integrity management across the content life cycle. That is, the creation and monitoring of checksums (hash values) would be more straightforward with a simple file that contained fewer content elements. Our AS-07 specification will use MXF elements that support file integrity management. We plan to cover this topic in a more comprehensive way in the near future.

Appendix A. The query to the lists (slightly abbreviated)

Subject: Embedding Associated Materials and Supplementary Metadata in AS-07 files

The AS-07 planning team has struck a dilemma and seeks advice and comments from interested persons. Here's the nub. We recognize that there are various ways to package related content items together. For example, there is the important and soon-to-appear SMPTE AXF standard (it _binds files_ together and provides some important features related to storage systems); other options include BagIt, METS, and others. But we have also heard from archivists who wish to _embed_ certain kinds of materials right _in the MXF/AS-07 file_ itself, encapsulated in the same essence wrapper.

The desire here is not to wrap together every possible related entity but rather to provide a mechanism to embed "stray" associated materials. This desire has arisen in memory institutions as they reformat old videotapes but it will also arise in the case of born digital content that is heading for long-term storage. What entities have archivists talked about embedding?

- Images of things on or in the old tape box: scribbled notes; documents of one kind or another, including scripts-on-paper; and 8x10 photoprints.
- Metadata (beyond what MXF demands) that provide good information about the content item, or that logs the result of the transfer process.
- Edit decision lists, maybe printed out on paper or in machine readable form.
- Scripts or other texts in machine-readable form ("the floppy you find in the tape box")
- For born digital items, a video trailer.

Our terminology for these entities is _Associated Materials_ and _Supplementary Metadata_. Anyone who has worked in a media archive will think of many more examples. These materials contribute to the completeness, comprehensibility, authenticity, or contextualization of the main content in the audiovisual streams. They are secondary to essences, however, and would not be reflected in the MXF file's Operational Pattern.

What's the dilemma? Some team members ("specifiers") want to (a) constrain what can embedded to well defined file types and (b) require decoders (understood here to mean the entire file-reading and playing system) to display the embedded element and/or create thumbnails (or icons for types of files) on the fly. One team member pointed out that the constrained list requires a certain performance level from decoders, and supports greater inoperability.

Meanwhile, other team members ("free-formers") want to (a) allow for almost anything non-viral to be embedded (shuddering at the thought of executables) and (b) only requiring decoders to report that they have an instance of "x" (as labeled by the file creator) and to hand over the bits. (The recipient will have to find a player, thank you!) This approach allows wider freedom for the tool makers and users to adapt this feature to specific environments.

If we go with the specifiers, we might constrain the permitted files along these lines:

- Bitmapped images: TIFF, JPEG, JPEG 2000, PNG, PDF
- Machine readable data and texts: TXT, XML, CSV, PDF

- Video (secondary, remember, not the real essence): UNCOMPRESSED, JPEG 2000, MPEG-2, MPEG-4
- Audio: BWF, MP3, AAC

If we go with the free-formers, the types are open-ended. But in both cases, we feel that file-makers need to tag items in the file, in the header and/or in a manifest that we will require: (i) file type extension, (ii) MIME type, (iii) version or profile information if relevant and if known, (iv) short note or prose description (optional), (v) size (extent in bytes), and (vi) location in the file.

As we continue to deliberate this topic, we would like to hear from you:

1. Would you embed Associated Materials and/or Supplementary Metadata in files destined for the archive? Does this seem like a helpful idea?
2. Do you vote with the "specifiers" to constrain the types of files to embed, or with the "free-formers" to leave it open?
3. What other comments would you offer to the AS-07 team?

Thanks and best wishes.

Appendix B. Summary of Discussion on the Listservs and Forum

- Files need rules in order to function so constraint is imperative.
- Embedding associated materials may not be the most efficient way to organize and access the data for downstream use. Instead, embed sparingly and even then, use pointers to external sources of information.
- Some feel that adding further complexity to the wrapper design is a bad idea. There are other options for linking data to create packages like METs or BagIt
- Some confusion of the roles of AXF and MXF – what does what?
- Voice of concern: will the complexity AS 07 limit its adoption in the global market, especially in developing countries?

Topic	**Contributor**	**Comment**	**Summary**
Alternatives to using embedded Associated Materials	University archivist	…..it is my belief that continuing development of linked open data and semantic web technology will gradually quench the desire to contain everything other than the bare essentials in a single file	Other alternatives are available to achieve the same goal – linked data, semantic web
Alternatives to using embedded Associated Materials	National library officer	I would embed the MXF in our METS schema, using the relations section to take care of Associated Materials and/or Supplementary Metadata	Other alternatives are available to achieve the same goal – METS
Alternatives to using embedded Associated Materials	Industry technical expert	…now the METS tag idea has me thinking again…	Other alternatives are available to achieve the same goal – METS
Complexity	Broadcast archivist	There are moves toward smaller files, like DPX, and toward clear and simple ways to prove the integrity if	Complexity is difficult to support

		each frame, as in checksums for each frame.	
Complexity	National archive technical director	I fully agree with [his] opinion. We've been promoting, and of course using, such a mechanism for years now and it proved to work flawlessly in most situations. In addition to tar files or Bag-it containers, also a properly implemented directory structure does the job.	Complexity is difficult to support
Complexity	Broadcast archivist	The future us about enhancing and securing the performance, not adding to the structure…	Complexity is difficult to support
Complexity	National library officer	…. Remember that the standards accepted in developed countries impacts on the capabilities of developing countries.	Complexity is difficult to support
Complexity	Industry technical expert	Given [his] scenario I would still advise on using a constrained compatible format for the essence instead of a vendor specific implementation, and solve the database problem by packaging the related files with Bag-it or tar or something similar, using a naming convention to relate the files to each other.	Complexity is difficult to support
Complexity	Broadcast archivist	Currently I'm thinking that it's best to keep the AS-07 format as simple as possible while preserving the full details of the essence (video, audio, timecode). Having all of these elements inside the MXF file is useful because MXF is able to handle the specialist details of these elements and provide the internal 'file system' needed to relate them in the time domain. To ensure that the full details of the essence are preserved items such as Dolby-E metadata probably need to be considered (I've not recently checked what's on your list already) e.g. as is currently being considered for AS-11.	Complexity is difficult to support
Packaging/bundling	University archivist	The specification should allow only enough embedded content to link the essence to external assets reliably or to	Bundling data not beneficial for long term -

		identify the essence in case of systemic failure…..allowing extensive embedding encourages those that are searching for the perfect container that we have already learned does not exist ….continuing development of linked open data and semantic web technology will gradually quench the desire to contain everything other than the bare essentials in a single file	Use embedding sparingly and to point to other external data sources only
Packaging/bundling	National archive technical expert	An additional argument for "pure" formats is the use of checksums. Generated at the beginning of archiving the file the checksum is the most reliable option to prove the authenticity later on. I wouldn't recommend to put anything into the archive file which eventually has to be altered later on. Descriptive metadata i. e with the time may need to be corrected or extra information has to be added. This would lead to a different checksum when calculated again. By the way: The checksum should be stored outside the archive file itself. This is just one point why I don't see it realistic to be able to keep all Information concerning the recording in one file.	Bundling makes updating more difficult
Packaging/bundling	Broadcast archivist	Trying to do everything in one file us no longer progressive thinking….	Bundling is not the direction the community is going
Packaging/bundling	National library officer	I would not embed Associated Materials and/or Supplementary Metadata in the MXF file	Bundling data not beneficial for long term
Packaging/bundling	Broadcast archivist	All other items such as videotape box images, programme scripts, etc are netiher 'specialist' A/V media items nor 'time dependent' and so can be handled by a standard IT 'file system', packaged 'loosely' (e.g. BagIt, METS, etc) or 'tightly' (e.g. TAR) or both, as required. If items are 'time dependent', as long as they do not need to reference the essence with sub-frame accuracy then the	Use MXF for main essence. Use other means for secondary material

		frame timecodes can be used to reference the essence. These sorts of items may be added to the archive over time and it could perhaps be problematic to update the large AS-07 file to add them. Also, even if a constrained list of file-types is allowed for inclusion in the AS-07 file people will later think of more things to include and there will be a temptation to expand the list of allowed items...	
To constrain or not	Industry technical expert	As [he] said, stay with the specifiers. Interoperability in an open system requires a limited and well-defined set of features, as otherwise cost of widespread compliance will be too high and won't happen. A closed system can have any private set of extensions.	Constrain
To constrain or not	National library officer	the perspective of a collection manager/curator in a progressive thinking, well funded organisation with a well developed collection management system (and in fact entering the third generation of development) and data storage and integrity capability, the right thinking approach (which is clearly right because all of us right thinking people agree) of constraining the types of file that can be embedded and ensuring that our collection management system and associated local/widely-accepted schema accepts it.	Constrain
To constrain or not	University archivist	…restriction is a good thing, at least initially, in order to get development and commercial support rolling…. Embedding players and executables seems like asking for trouble--not just because the players will become obsolete and unsupportable….	Constrain
To constrain or not	National library officer	What we need and want is compatibility, and I do believe that is best served by strong constraint.	Constrain
To constrain or not	Broadcast	Files are rules and operations that orchestrate a performance that does something with data. Keeping the	Constrain

	archivist	performance simple, granular and verifiable should be the main concerns	
To constrain or not	Industry technical expert	MXF AS-07 is a constraint with intent…... AS-07 is intended to inform media players and media engines how to process the file, so I must align myself with the "specifiers" who recognize the limits of information that can be passed to those automated processors.	Constrain