BROADCAST ARCHIVES: BETWEEN PRODUCTIVITY AND PRESERVATION

Jean-Christophe Kummer, NOA Audio Solutions, Austria
Peter Kuhnle, NOA Audio Solutions, Austria
Sebastian Gabler, NOA Audio Solutions, Austria

1. Abstract

File-based digital archiving has been in use in the AV sector for the past 20 years. Recommendations such as IASA-TC 04 for audio have helped give archivists a high awareness of quality when digitising archive material. The Open Archival Information System reference model provides assistance with the organisation and layout of long-term archiving. Specific requirements regarding the usability of an archive, however, present broadcast archives with more sophisticated requirements. In this article, the authors consider different aspects of the appetite for AV broadcast archives and the suitability of data formats in the AV sector from historical, current, and forwarding-looking perspectives.

2. What is the purpose of a broadcast archive?

Asking this question quickly sheds light on the different requirements of archivists and producers within a broadcasting company. The range of responses here spans from "conserving cultural heritage" to "a backend for transferring out of the production system."

The purpose of a conventional archive is typically the collection, description, and preservation of physical carriers. The user has no direct access to the media content, and can only view its metadata. In these archives, content is only available once the corresponding carrier is played back.

One of the main reasons for digitising audiovisual inventories is most often related to the laboriousness of this setup.

2.1 The archivist's perspective

For traditional archivists, the prevention of amnesia — or the preservation of archival material — is key.

Archivists also highlight the need for a media library, to enable the cataloguing, searchability of, and access to the material. Whereas the main requirement in preservation is maintaining as exact a copy of the original as possible (formal cataloguing), when it comes to the process of actually cataloguing the material, it is the indexing which is key — that is to say, description of the content's features and its availability. To this end, public archives generally use standardised archiving rules[29] with their associated data schemas. Due to their level of detail, only significantly slimmed-down versions of these schemas can be used in broadcasting archives. Because of the file format, the availability of the media now becomes merely a technical and logistical problem, since the laborious process of creating or lending copies no longer applies.

2.2 Producers' needs

On the other side of the catalogue we have the producer who, given the task at hand, wants to use the archive material as efficiently as possible in new productions. Catalogue searches are content-oriented and so, from an archiving perspective, both the logical, original context, as well as the original format of the archived object, are more of a hindrance than a help. Since production generally only requires small segments in a specific file format, it makes more sense

29 Typical set of rules: MARC21 in the Anglo-Saxon world, www.loc.gov/marc/; MAB2 used to be prevalent in German-speaking countries, but was finally superseded in 2013.

if they can be retrieved from the catalogue directly as required. Direct delivery from production systems with the associated metadata in a standardised format helps to minimise cost. Heritage archive material should also be made available in its restored version in order to keep the production process as lean and fast as possible. Long processing times should also be ruled out as much as quality-reducing format conversions.

2.3 Potential solutions

One approach is a logical division into an archiving area and a cataloguing area (mediatec or media library).

While digital originals or their equivalents are managed in the archiving area after careful digitisation — observing the recommendations of the IASA-TC 04[30] wherever possible — the cataloguing area contains reorganised entities, derived from factual descriptions, so-called objects of the 'segment' type, e.g., a documentary. Corresponding media can be linked several times and in sections with the metadata, meaning that physical copies or copied sections can be avoided (Figure 1).

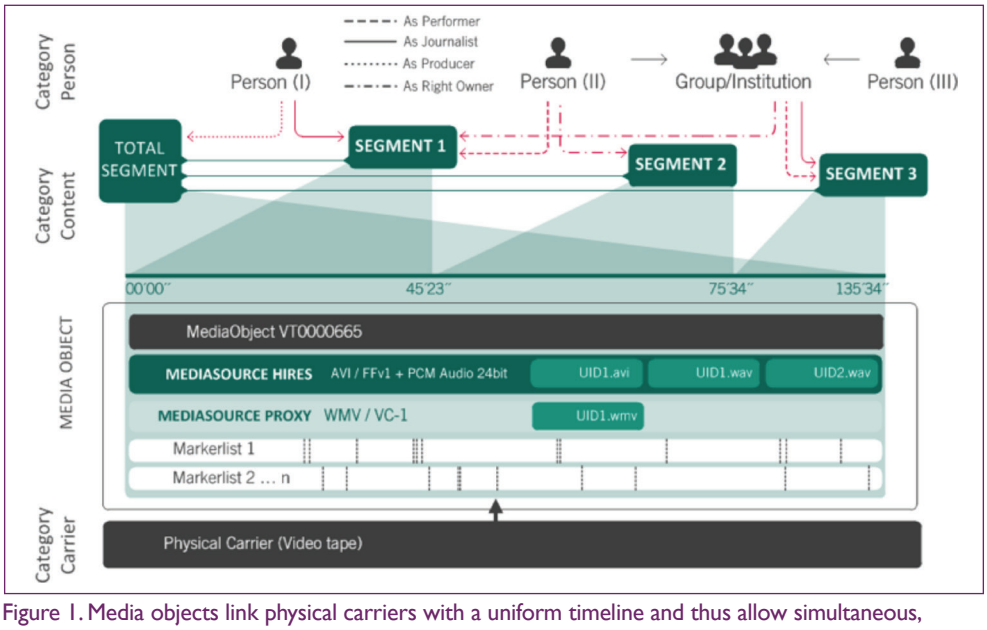

Figure 1. Media objects link physical carriers with a uniform timeline and thus allow simultaneous, consistent access to different codings and segments of the same content. People are then linked by semantic role (author, producer) to the segments (system: NOA mediARC).

Media management should also hold content in several parallel formats so that restored copies, digital originals, and working copies can be stored together. This means that the parallel, restored copy of archive material can then be made available in the catalogue if the quality of the original version renders it unsuitable for publication. Likewise, it is also necessary to create a filing structure for fragments of recordings, which cannot be digitised in a coherent manner. Should the need for a restored copy arise during cataloguing — or the quality control checks carried out when adding new content — the system should be able to control the relevant processes to achieve this. It is often possible to restore source material automatically with few qualitative faults; to this end, 'official' external modules should be pluggable so that it is always

30 http://www.iasa-web.org/tc04/audio-preservation

possible to work with state-of-the-art technology. By contrast, initiation and control of semi-automatic processes, which also incorporate external editing software, help ease the handling of more seriously damaged material.

In order to provide timely access to sections of a catalogue object in the desired target format, the cataloguing system should always include an additional working copy in a format that can be processed quickly and universally, if the original format does not allow this. The Open Archival Information System, known as OAIS[31], makes the distinction here between the Dissemination Information Package (DIP), the output for the consumer, and the Archive Information Package (AIP), which constitutes a consistent, easily readable description of the archived object (Figure 2).

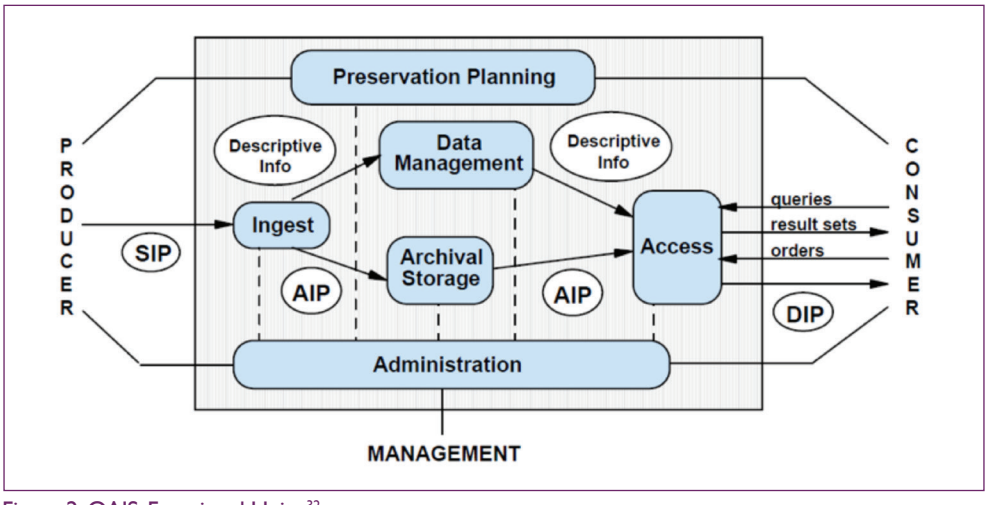

Figure 2. OAIS: Functional Units.[32]

When requesting media, the user has the choice of requesting an original copy or a clip in one of the supported target formats. It should also be possible to deliver metadata in parallel. Freely configurable modules should be available, given the variety of formats necessary (Figure 3).

31 OAIS is a reference model for a dynamic, extensive archiving information system and is rooted in ISO Standard 14721:2012.

32 Picture source: http://en.wikipedia.org/wiki/Open_Archival_Information_System.

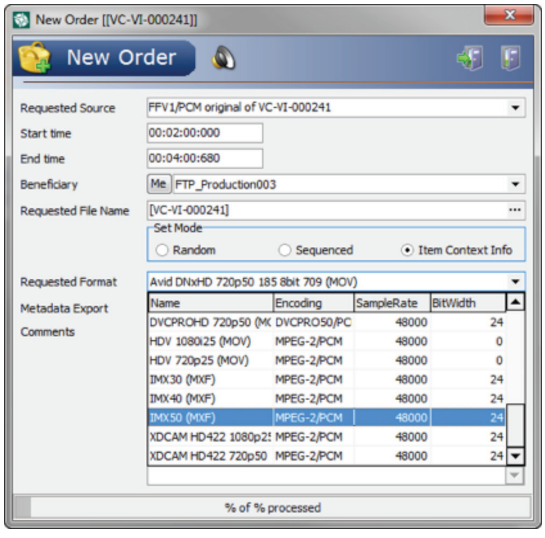

Figure 3. Placing a cart order for a 2-minute segment in a specific target format, querying items stored in various original archives (here FFv1[33]). The beneficiary in this instance is a production system's FTP server gateway (system: NOA mediARC).

Linear archival formats, from which segments can be exported, have long been in existence in the audio sector. Nevertheless, an attempt shall be made to look critically at a variety of different archiving formats in the audio as well as video sector.

3. Fifteen years of BWF in audio archiving — a success story?

As an audio archival format, the Broadcast Wave File[34] certainly has its advantages. Years of use in archiving raise the question: would we go down the same route again?

Looking back on those formats that could be read fifteen years ago and those that are readable even today, it is clear that some are still in use. Whether *.txt, *.jpg, *.pdf, or *.wav are viewed as suitable, these formats do have one thing in common: they are easy to open.

It is probably safe to assume that an archive based on formats for which there are hundreds of programs today will still find some form of code that can read them in 30 years' time.

From where does the idea of a specific archive format stem?

The archiving world tends to connect all metadata of a technical as well as content-related nature to the archive object, presumably in order to also preserve, in the digital world, the physical idea of the carrier on which the metadata is located. In the OAIS specification, the AIP provides for the creation of clearly readable archive objects, something that enables archived media to be read and interpreted, even when the necessary devices and applications are no longer available. Yet it does not outline a single, descriptive file that can be used for archiving purposes. Looking back to the early file-based audio world, it is clear that acceptable solutions for workflows and archives were established in the production and archiving environment — primarily as a result of the 'tapeless' file sizes — a good ten years before the video world.

33 FFv1: mathematically lossless Open Source Video codec, reference: www1.mplayerhq.hu/~michael/ffv1.html; also www.digitalpreservation.gov/formats/fdd/fdd000349.shtml.

34 http://www.digitalpreservation.gov/formats/fdd/fdd000356.shtml.

In the 1990s, it emerged that *.wav, as a basic RIFF format, would be suitable as an audio archive format, both because it uses linear encoding and also because it can be played back using the integrated player on any Microsoft operating system, without the need for a license or any significant additional hardware costs.

Consider the background to this realization. Very few professionals worked with *.wav on a regular basis, since the majority of professional interfaces in the 1990s were still using Apple-centric solutions (SoundDesigner, Sonic Solutions, etc.), which worked with SDII or AIFF formats.

The willingness to consider a format known during this period as 8-bit system sound (for which *.wav was being used primarily) as a suitable archive format, albeit in a slightly modified version (24 bit/48 kHz), was therefore the product of altruistic thinking and an example of significant foresight by the archiving world and industry. *At the time, the main rationale for this was the widespread use of playback systems.*

Most conscientious producers took to redundantly storing all technical metadata in an index in the archive system, so that the archive data could still be filed at a later date in whatever header metadata structure would become available. Despite this, it took a long time for a clear specification to emerge.

In hindsight, does the BWF specification still make sense today?

Today, a few things would probably be done differently:

- A file format that accepts only ASCII characters[35] in the descriptive chunks would, today, be considered unsuitable: character sets such as those based on the Latin alphabet, including even German special characters (ö, ä and ü, for example) would otherwise be completely stripped out. Most people would now choose to follow at least an ANSI or Unicode implementation instead.
- The fact that coding history has been implemented in a variety of ways means that semantic, automated processing is often unachievable.[36]
- The interpretation of the peak file as a component of the header may, on occasion, be understood by a system. At other times it may be ignored by others, or even overwritten with a proprietary version.[37]
- The proliferation of RIFF chunks has led, at best, to a correct selection characteristic in the production system,[38] but is still a classic case of a specification's 'vendor lock-in.'

In 2005, it was already clear that MAM systems could not make much use of the information embedded in the header. A workaround was created by requiring systems to produce a 'tandem of files.' This consisted of textual information of whatever kind (XML, TXT – potentially following a standard such as METS[39]), most often paired with the audio file of the same name.

Some manufacturers certainly made an effort to export the chunk information (David, Blue Order[40], and NOA, among others), but ultimately only time will tell whether the information lying dormant in the headers will be of relevance in 30 years' time, and whether code to interpret it will still be available.

35 http://tech.ebu.ch/docs/tech/tech3285.pdf, ASCII coding of chunk information.
36 https://tech.ebu.ch/docs/tech/tech3285s2.pdf, broad definition of "CodingHistory".
37 https://tech.ebu.ch/docs/tech/tech3285s3.pdf.
38 https://tech.ebu.ch/docs/tech/tech3285s2.pdf, FileSecurityReport and FileSecurityWave are only available from a manufacturer (VendorLock).
39 http://www.loc.gov/standards/mets/.
40 BlueOrder, Kaiserslauten was taken over by Avid in January 2010.

If nothing else, the work involved in parsing headers (in comparison to text-based files) requires special tools today, which are often not justified given the effort involved.

The technical utility of embedding descriptive metadata is demonstrated from a purely practical point of view: currently, there are many large-scale digitisation projects being carried out in both Europe and other areas of the world, some already completed, in which material is being digitised retrospectively.

Digitisation itself usually only accounts for 10% of the overall time needed for these projects, with the remainder spent managing metadata, clarifying rights, and handling carriers.

Commercial broadcasting companies' projects are set up in such a way (see the Swedish broadcaster SRF, Slovenian RTV and Croatian HRT) as to achieve a certain result in a limited time.

In the case of SRF, the task was to digitise 220,000 hours of ¼-inch material, DAT, and CDs within three years (Figure 4). This was successfully carried out within the target period, with the appropriate quality control, at eight special recording stations using a shift system.

Figure 4a to 4f. Clockwise from top left to bottom left: SRF recording studios (audio and video), 1-inch MAZ machines, videotape archive, tape/Senkel, and CD archive collections.

Adding the 10-fold larger job of correct metadata research to the impressive effort of digitisation would have meant that even filling the BWF 'description' field with a correct, comprehensive working title (before writing the BWF) would probably not have been affordable.

Since the system used relies on the premise of an extremely robust relationship between database record and media object, subsequent socialisation of content becomes possible.

As a result, this information is unlikely to end up in the header of the already digitised BWF file, which may even be exported to LTO tape. Instead, it is more likely to be kept in an easily searchable, non-redundant index of a low-maintenance, migratable, and readily available standard database that offers multiple ways to access the data, other than through the manufacturers' interfaces alone.

Without wanting to sound critical, it seems logical to conclude that for audio archiving at least, it is not absolutely necessary to store descriptive metadata in the header of an archived object, but rather to keep only a rudimentary archiving number and technical metadata within the file, provided it is possible to interpret and process this data automatically in a sufficient number of different systems.

In doing so, it might be easier to migrate separately stored metadata from an easily readable format into the next archive format. However, it seems essential that the RIFF file as a widespread codec in a WAV container as a linear essence best suits the requirement of essence storage.

As elegantly as archiving in the audio sector can be solved using WAV/BWF and the optional TandemFile, the range of choices for the world of video is just as diverse.

4. Making video archive copies for long-term archiving — lossless or lossy?

Broadcast archives are based on the premise that current production formats, even those that use data-reducing codecs, need only to be exported to tape-based long-term storage to fulfill the criteria of 'archiving' content. But are there any other approaches?

4.1 Codecs

Before diving into the accompanying metadata and how it can be stored, the crucial question of which codec to use still remains. In the world of video, a codec's suitability for use in production often dictates the chosen archive format, especially for public broadcasting companies who, by their very nature, focus primarily on the level of productivity of the archive.

Twenty years of experience in audio digitisation have shown that retrospective digitisation archive projects, which were implemented on the basis of psycho-acoustic data-reducing formats (cf. MP1L2 archive in the mid 1990s), are of practically no real significance today. Digitisation projects were even repeated in linear formats provided the funds could be found. Looking at the effect of multiple codings of data-reduced formats within a video segment, it can be seen that by cascading several encoding steps, there is a successive loss of source information (Figure 5).

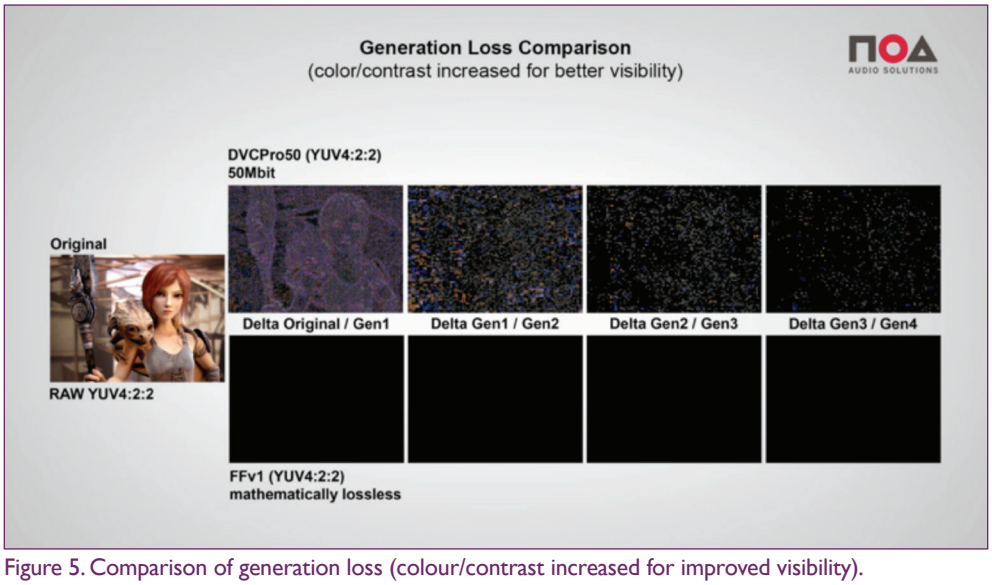

Figure 5. Comparison of generation loss (colour/contrast increased for improved visibility). The sequences of dark images display the respective differential signal to the precursor signal. The multiple coding has absolutely no effect in the case of mathematically lossless coding (lower row).

4.2 Linear video? In the production archive? For SD material?

The cost of the memory required, the burden on networks, and the unsuitability of lossless encoding in the production system are obvious reasons for avoiding such a course of action.

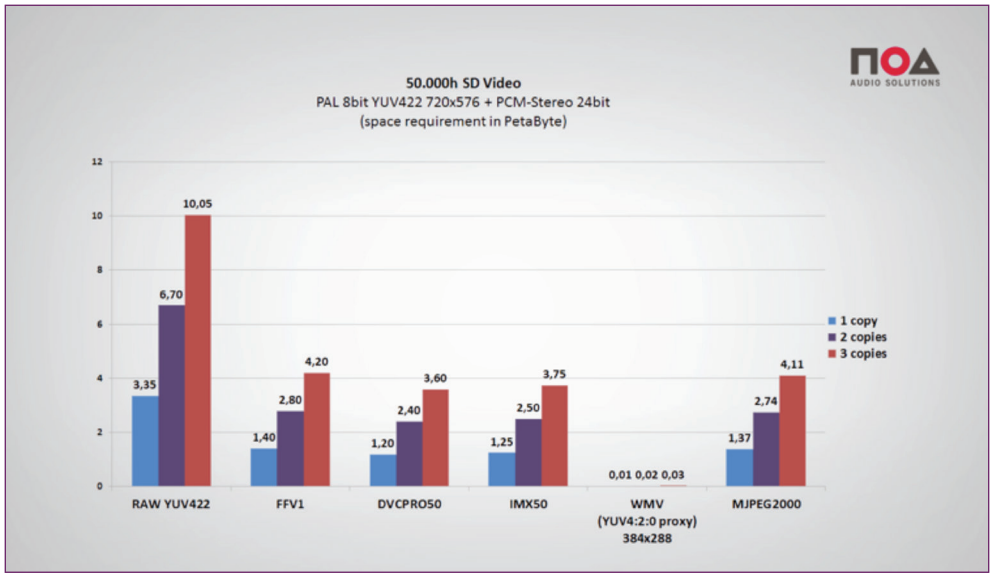

Figure 6. Comparison of memory requirements for 50,000 hours of SD video in petabytes (PB) (PAL format: 8 bit, YUV 422, 720×576 and PCM 24 bit stereo).

Considering the memory requirements of linear versus typical compressed material, the following applies (Figure 6). As can be seen, mathematically lossless formats such as FFv1.3 or MJ2K[41] (also called MJPEG-2000) result in only marginally larger files than data-reducing production formats such as DVCpro50 or IMX50.

Although MJ2K in particular was promoted within the Library of Congress, FFv1.3 has since emerged as a codec alternative. Filed in an AVI container,[42] it meets archive requirements for simplicity and ease of readability with some minor known limitations.

Regardless, three copies of 50,000 hours of SD material would take up the rather impressive total size of around four petabytes: a cost factor which, in any case, still needs to be considered. Nevertheless, in times of tight budgets it seems a little excessive to follow the purist idea of storing archives in a completely linear way. It also seems avoidable, given that doing so increases the overall cost by 100%.

If, therefore, there is little difference between lossless and lossy memory requirements, the question of the amount of time required for conversion of the archive format to a production-ready version remains (Figure 7). The ability to convert FFv1.3 to DVCpro50 at almost eight times the rate of MJ2K could be one of the reasons — alongside ease of use — why FFv1 has gained importance.[43]

41 MJ2K – Samma Profile: Beschreibung: http://www.digitalpreservation.gov/formats/fdd/fdd000271.shtml.
42 http://www.digitalpreservation.gov/formats/fdd/fdd000349.shtml.
43 FFv1 implementations: www.Archivematica.org; www.digitalpreservation.gov/formats/fdd/fdd000343.shtml; Open Source Digitisation of Österreichische Mediathek; City of Vancouver Archives; NOA MediaButler.

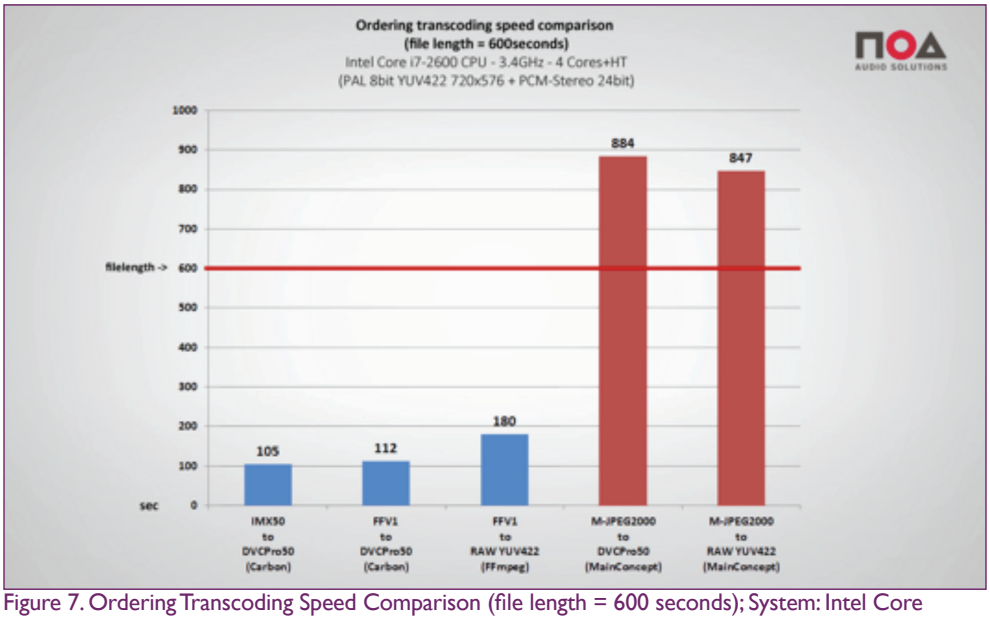

Figure 7. Ordering Transcoding Speed Comparison (file length = 600 seconds); System: Intel Core i/-2600 CPU, 3.4 GHz, 4 Cores+HT; PAL-Format: 8 bit, YUV 720 576 + PCM-Stereo 24 bit.

In addition to this, it has become apparent that converting between two lossy formats takes almost the same amount of time, without resulting in any additional benefits. Strictly speaking, it should be added that the data is already contained in MXF wrappers[44] before the whole file is written.

4.3 What do others do with their inventories?

Broadly speaking, there are two main trends. Large national archives (e.g., the U.S. Library of Congress[45]) with generous budgets can afford to use three memory formats at the same time, with the corresponding cost structure:

- linear or mathematically lossless formats, e.g., linear.mov, ffv1.avi, j2k.mxf
- production formats, e.g., MXF D10, DVCpro50
- proxy formats, e.g., WMV, H264

By contrast, broadcasting companies only use:

- production formats, e.g., MXF D10, DVCpro50
- proxy formats, e.g., WMV, H264

4.4 Are there any alternatives?

As a result of an EBU study[46] conducted in 2010, it was established that delivery from long-term archiving systems to production can take, on average, up to ten minutes.

44 MXF Smpte Spezifikation: Op1a: SMPTE 378M.
45 Report Carl Fleischhauer IASA 2013 Vilnius.
46 https://tech.ebu.ch/docs/techreports/tr006.pdf.

A closer inspection of the content requested from the archive shows that requests are made primarily for segments which need to be converted, where possible, into a different target format.

So why not consider the following:

- mathematically lossless formats, e.g., linear.mov, ffv1.avi, j2k.mxf, which actually fulfill archiving requirements and have tried and tested re-transcoding paths
- proxy formats, e.g., H.264

The effective advantage of MXF-based archiving files lies in the more reliable partial availability, which allows processing to begin before the transfer is complete. This, in principle, sways the balance in favour of MXF-wrapped MJ2K files or, if the lossy nature of the coding can be disregarded, generally only for lossy MXF files.

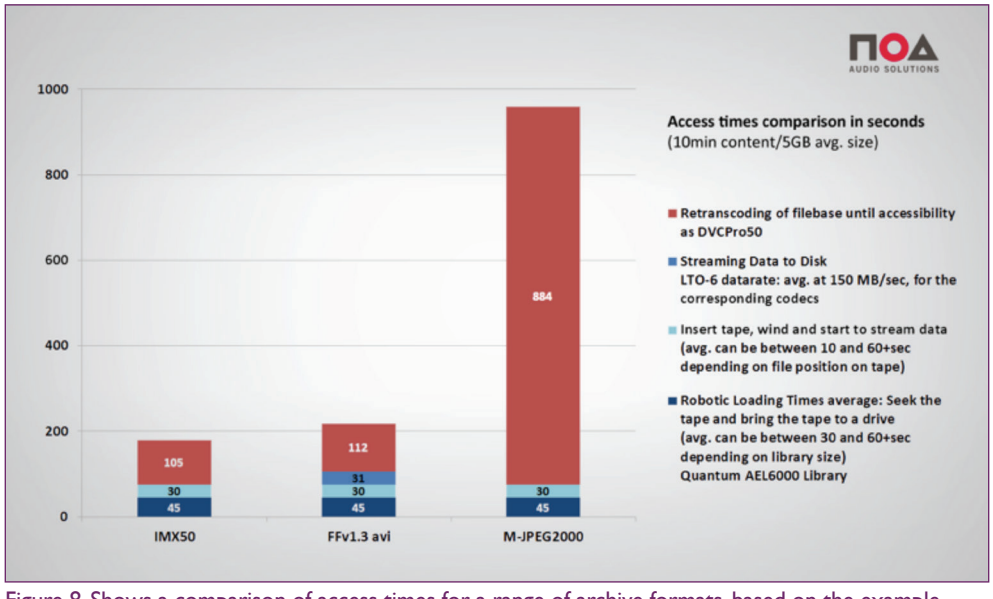

Figure 8. Shows a comparison of access times for a range of archive formats, based on the example of SD archiving; in this case, the transfer of a 10-minute video source SD file from an LTO 6 tape to a target production format (here DVCpro50).

If only allowing for lossless codecs from an archiving perspective, then the complexity of the MJ2K codec takes its toll when recording into or writing out from the archiving system. Also, a direct comparison of FFv1 and IMX50 as archiving formats is not entirely at the expense of the lossless compressed format when delivering to production (primarily with respect to the statistical means of robotics and tape-loading times). The economic savings in storage space are, however, considerable when taking linear versus lossless compressed video into account.

This means that today, production systems can be harvested from the lossless archive with DVCpro25 files — and other parallel systems with MXF OP1a-, H.264 or streaming files — which are then able to be exported, on demand, in the future.

Above all, what remains is the realisation that a sustainable archiving decision will have to be made again in ten years' time, since the life cycle of a production system tends to be between 3 and 5 years, and the production formats prevalent at that time may well be different to those in use now.

4.5 Wrappers

Video codecs become readable only when used in conjunction with a wrapper. In the same 2010 EBU study, broadcasters emphasised the serious nature of the problem of interoperability, giving it a score of 93 out of a maximum of 100. However, an IRT study[47] into the compatibility of manufacturers' MXFOP1a implementations in the production sector showed that improvements had been made, at least within the field of data reduction.

On the whole, the archiving sector seems to have the choice between several complex MXF versions for SD, provided that they are mathematically lossless, be it the MXF Samma profile MJ2K, AXF files,[48] or if one were to wait for the adoption of the MXF AS 07 specification.[49]

It seems inevitable that there will be some interoperability problems with formats that are not yet well-established (see previous commentary regarding experiences with audio). For this reason, some (in particular public video archives in the US) continue to opt for pure linear formats[50] and their corresponding tandem files, packaged using BagIt,[51] and are therefore forced to accept the 2.5-fold memory requirements.

The same cannot be said for RIFF-based AVI files, which can also be played back on any Windows computer. With ffdshow[52] or LAV filters,[53] the open source community also provides suitable packages for system-level decoding. They support codecs, aspect ratio, and multi-channel audio — all license-free and without the need for special hardware.

"Core" technical metadata is embedded in this wrapper. The storage of additional necessary technical metadata is carried out through a tandem XML file.

The conversion of archive material into a target format is a task that can be scripted easily with ffmpeg[54] or achieved using industrial encoders such as NOA MediaButler[55] or Harmonic Promedia Carbon[56] (previously Rhozet), which are able to read system codecs.

Since, according to the OAIS guidelines, production systems should never have access to originals and only intermediary instances should be used to produce the (broadcast-compatible) export formats, the Material Exchange Format (MXF) demonstrates its value once again: as a Dissemination Information Package (DIP), it is created from the archive format (FFv1 in avi in this instance) in line with demand, and the MXF then satisfies all transmission chain stability requirements with characteristics of a production format that can be processed in a swift and robust way.
If, in the next ten years, the archiving world manages to agree on a lossless video standard which also satisfies broadcast archive requirements, nothing stands in the way of lossless conversion from FFv1.avi into this target format using simple open source tools. FFv1 can be considered the perfect interim format that helps archivists and broadcasters in their decision.

Regardless, the operating archive will be able to migrate its archive data via simple scripting without the worry of vendor lock-in or help from the initial vendor.

47 http://mxf.irt.de/activities/2013-11-MXF-Plug Fest-ExecutiveManagementReport.pdf.
48 AXF has become a SMPTE standard, mostly driven by FrontPorch – now Oracle.
49 MXF AS-07 is developed within www.amwa.tv.
50 "*What Should We Do Today: Toward an Interim-Master for the Preservation of Digital Audiovisual Materials*" George Blood, Annual Conference 2011, Austin, TX, Association of Moving Image Archivists.
51 http://en.wikipedia.org/wiki/BagIt.
52 http://ffdshow-tryout.sourceforge.net/.
53 https://code.google.com/p/lavfilters/.
54 http://www.ffmpeg.org/.
55 http://www.noa-audio.com/products/actline/mediabutler/.
56 http://www.harmonicinc.com/product/promedia-carbon.

5. Conclusion

Decisions about the logical organisation of information and format issues in broadcast archives should, under no circumstances, be made from a production point of view alone. Requirements for archiving over a 50-year period differ greatly from the interests of production departments, which have to deliver results in much shorter periods of time. Appropriate technologies, which can generate linear or mathematically lossless archive formats, are provided both by the open source community as well as the industry. This development must be taken seriously because it is a pragmatic one.

A system that provides the archiving structure and that demonstrates organisational abstraction with respect to production in accordance with OAIS, while at the same time focusing on open standards and easily migratable media content, has a promising future and seems set to fulfill archiving requirements.