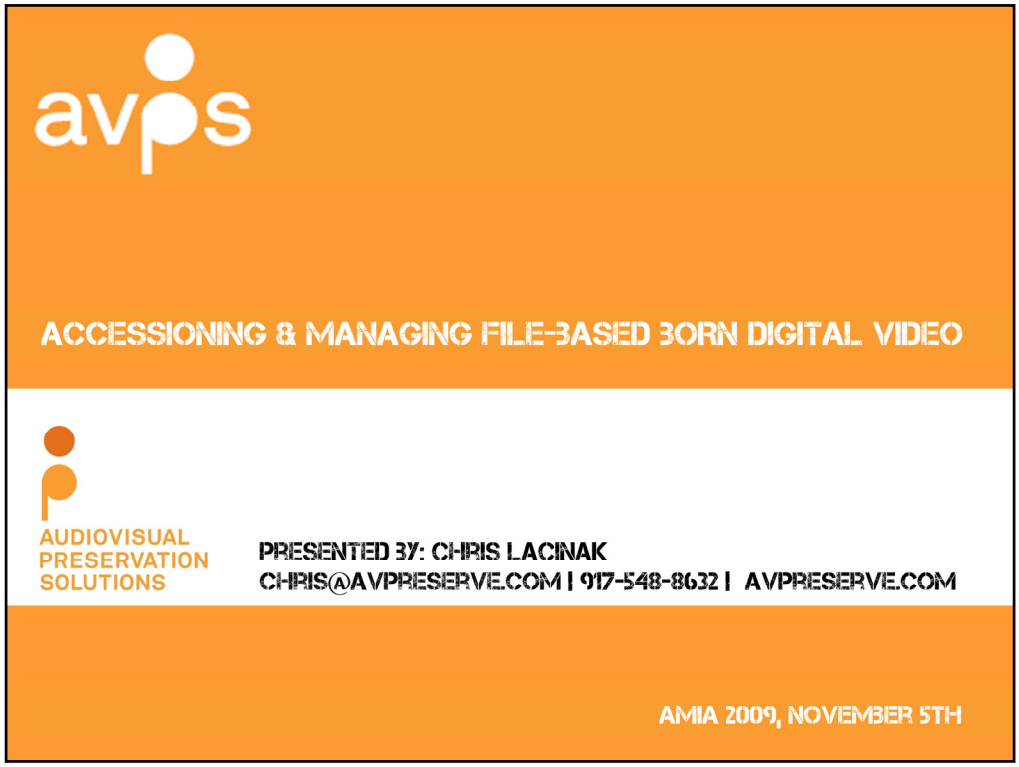
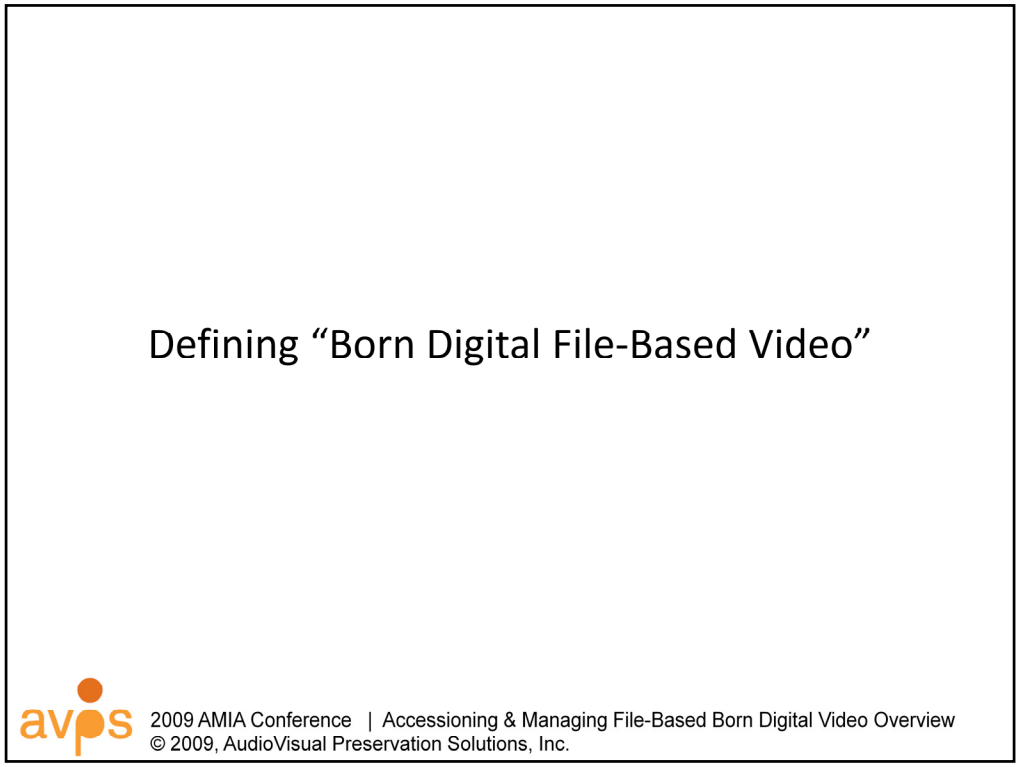
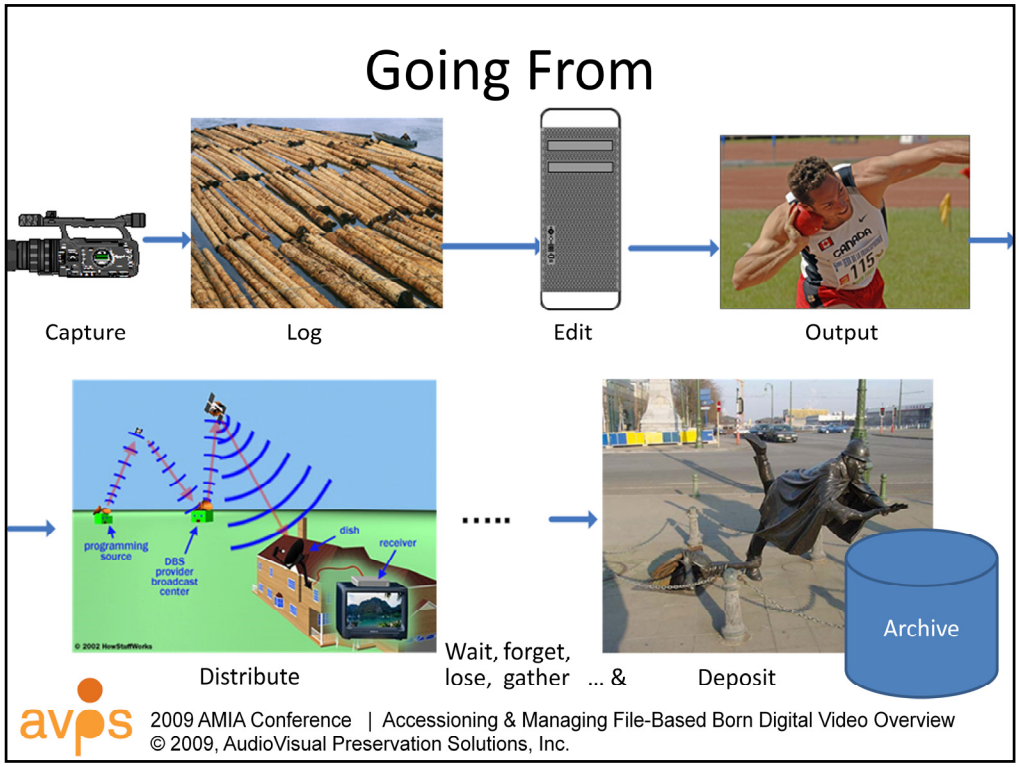


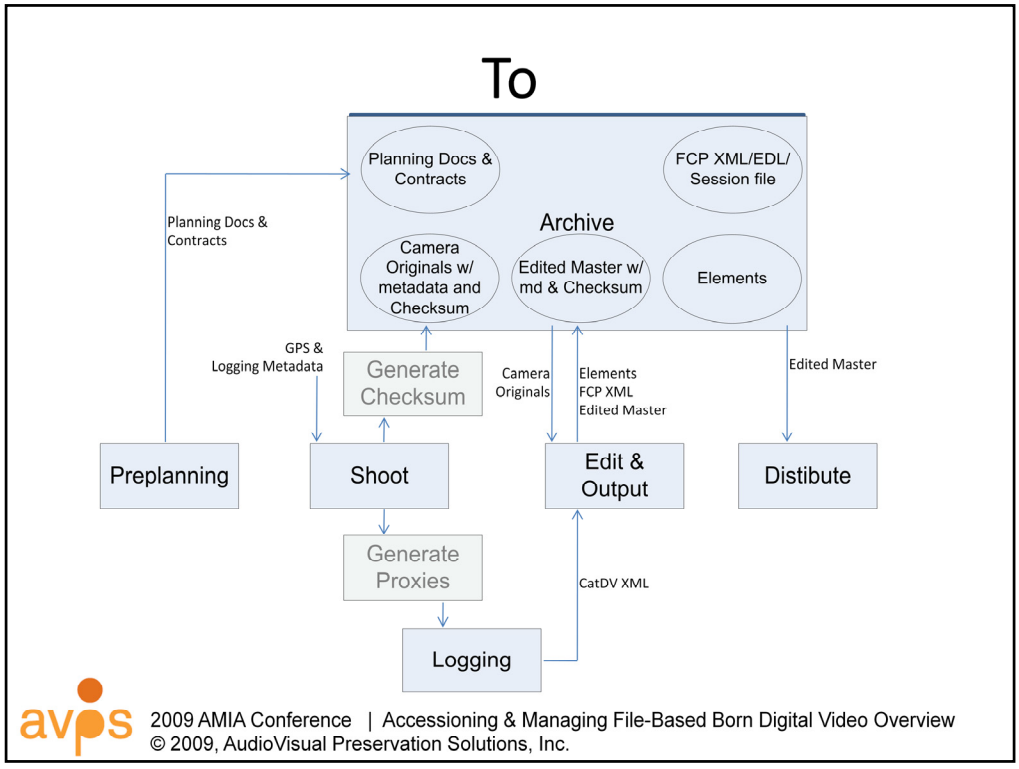
This presentation was given at the 2009 Conference for the Association of Moving Image Archivists in St. Louis, Missouri on November 5th. The session also included presentations from Grace Lile of Witness, Brian Hoffman of New York University and Dirk Van Dall of Broadway Video. Their presentations are not included here.



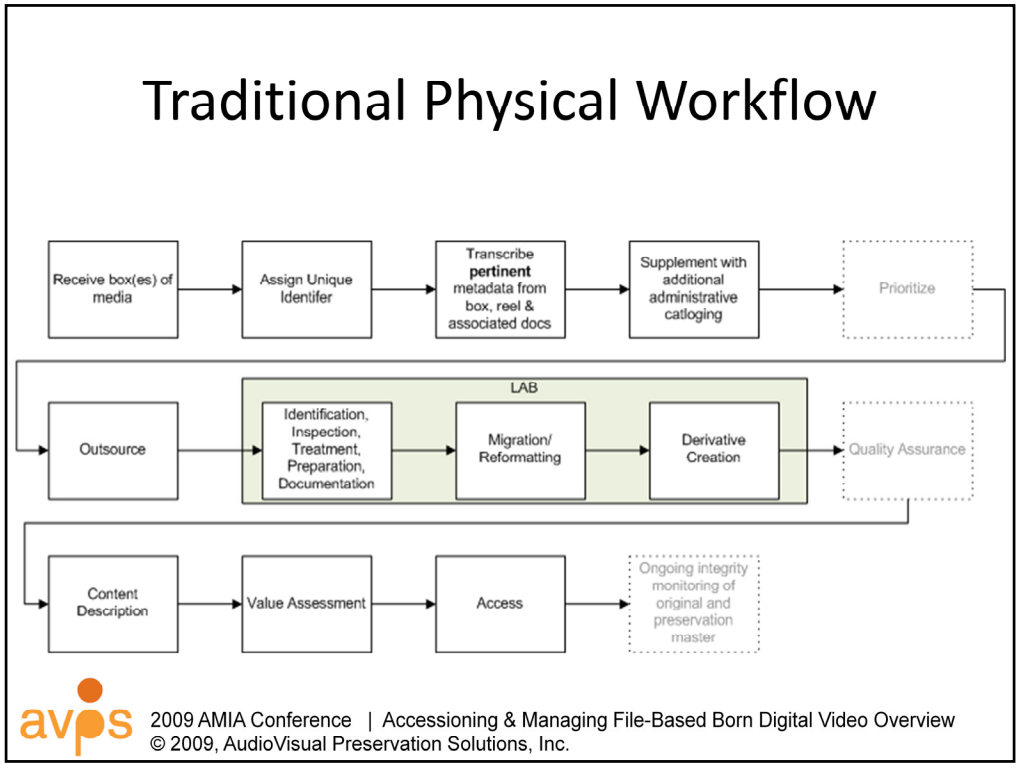
To define more specifically, what we are talking about in this session: Born Digital File-Based Video is a term used to describe video that is recorded digitally at its point of creation. Popular higher end examples of this include HDV wrapped .mov, DVC Pro wrapped .mxf, or R3D (RED) files. A popular consumer example of this is H.264 wrapped .mov files. To be clear, born digital file-based video *does not* include Digital Betacam tapes or files that have been digitized from analog media such as U-matic or Betacam SP.



Last year I talked about preservation oriented production and moving from this process, where content ends up in the archive in a haphazard way…



…to this process where the Archive is integral to the production process.
This year we are approaching the topic from the perspective of accessioning. While we still wish for the archive-as-central philosophy from both perspectives, on the whole, current workflows have not realized this as of yet, and many accessioning circumstances simply dictate alternative approaches. It's also important to recognize that the assumption in this diagram is that there is "a production". Whereas video collections of the past took extraordinary effort to create. Today video is in almost every personal device and is created with ease. **On a daily basis video is coming closer to being as much a part of the fabric of information creation and exchange as text.** The nature of video "collections" changes dramatically with this new reality and the types of video content being accessioned broadens significantly.

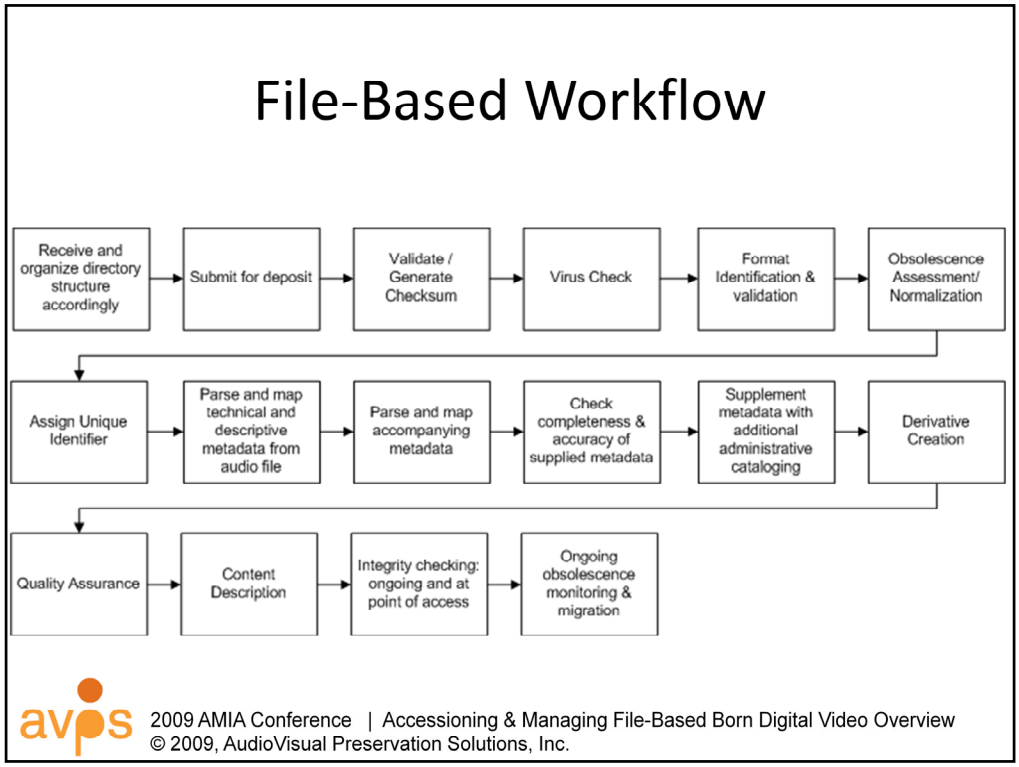


While legacy physical media is not our subject, *looking* at traditional physical workflows as a point of reference is a useful exercise.

In the physical domain many of the activities which enable intellectual control are back-loaded, not being performed until migration occurs.

The grayed out boxes are areas where I felt it remiss not to add as a point of reference, but too generous to imply that these things are *typically* done. And while the timeline is not represented here, these activities usually take place over a very long time span in the physical domain.

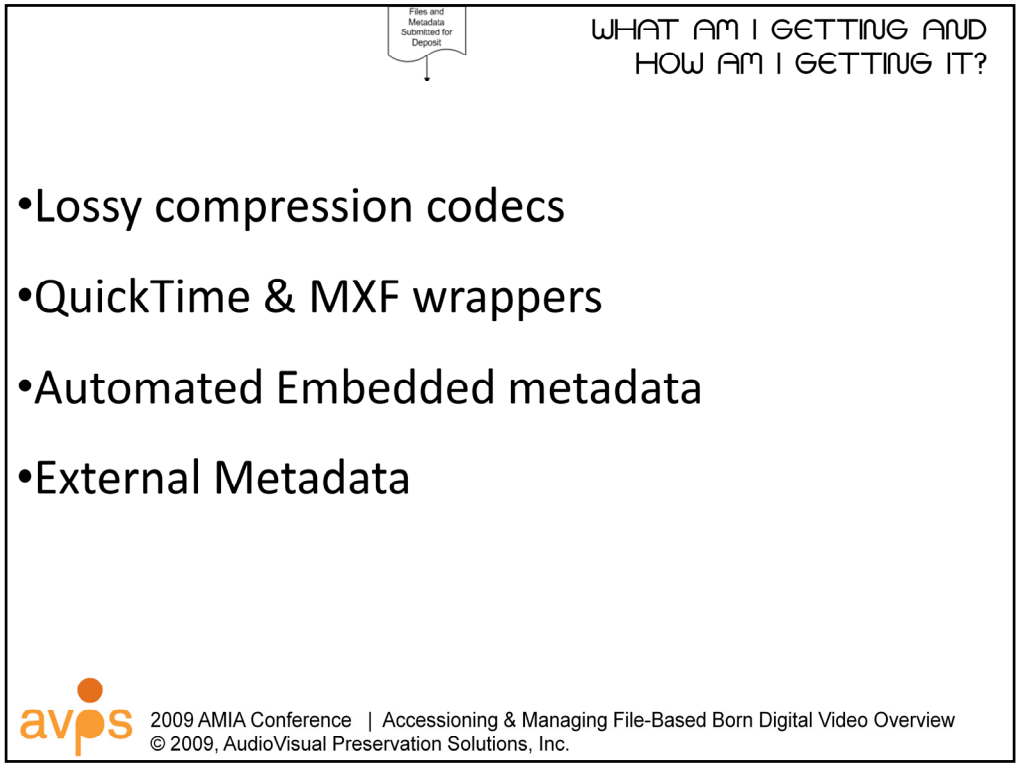
This lack of intellectual control prior to migration has been crippling to many archives. It has led, and continues to lead to the inability to effectively prioritize, budget and allocate resources. It disallows active management and performance of the OAIS functional entities.



With the move to the file-based domain, and the imperative of an actively managed environment, a change is not only required - but possible due to the availability of a new toolset.

In contrast to the long timeline necessary for the physical workflow, the timeline associated with this file-based workflow could almost be considered negligible.

Unfortunately, the general workflow shown above is not a typical one as of yet. It is practiced by relatively few. Fortunately, models and best practices are evolving. Today I want to go over some salient aspects of this workflow to identify challenges and increase understanding. So Let's start at the beginning.



What am I getting? The general trends are…

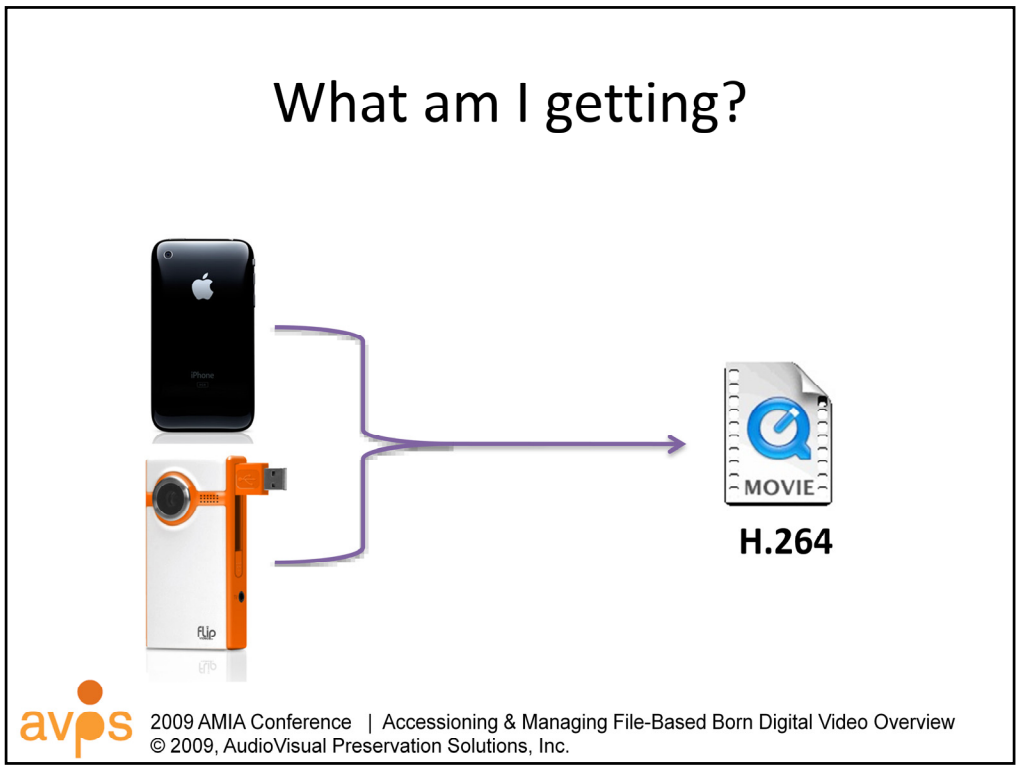
Lossy Compression Codecs: on the consumer end we are seeing H.264 or AVC codecs being widely adopted. While prosumer and professional cameras are using DV, HDV and Long GOP MPEG2. In the case of one very high end camera which you'll see shortly, it is using its own proprietary codec.

QuickTime and MXF wrappers: tend to be the wrappers of choice with AVI as a lesser used alternative.

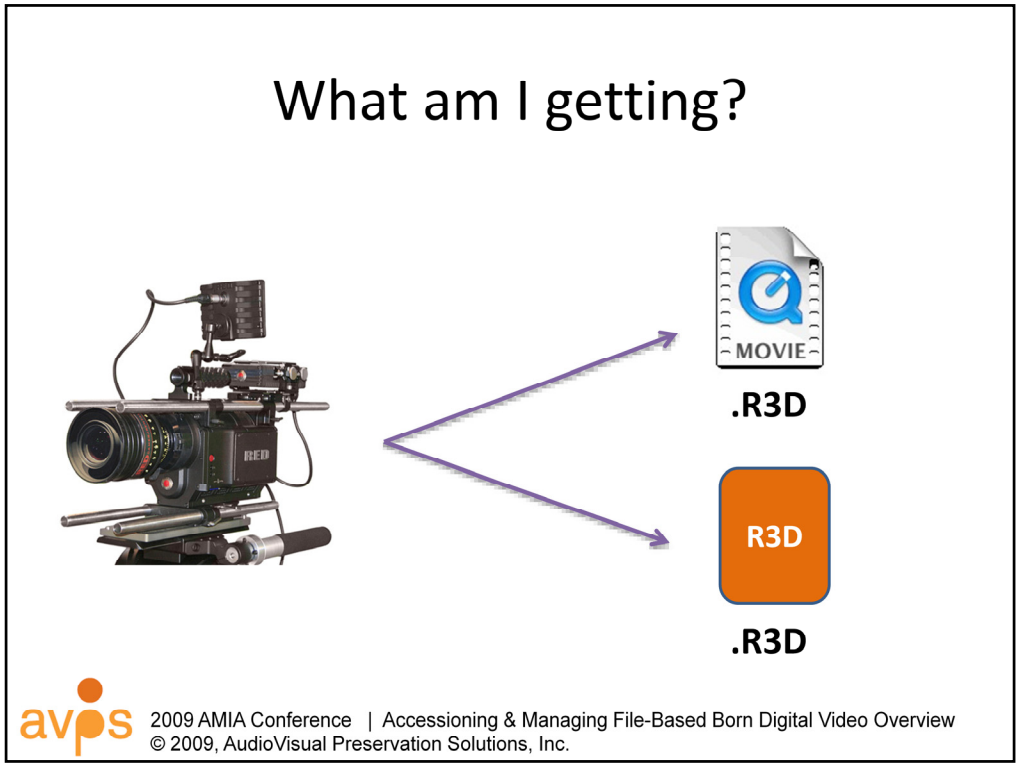
Automated Embedded Metadata: Cameras and recording devices are embedding GPS data, camera settings, and other technical and administrative metadata into the files themselves. Software such as QuickTime and Adobe Premiere enable extensive metadata embedding.

External Metadata: Recording devices are better supporting workflows and are more metadata aware. They are generating EDL (edit decision list) type data and they are enabling metadata entry in, and export from the recording device (camera/HD Recorder). Editing software such as Final Cut Pro are able to generate XML documents for interchange. This all equates to structural and technical metadata documents as part of the deliverables to an archive. Other inherent external metadata may also be inferred from the directory structure of the content being acquired.

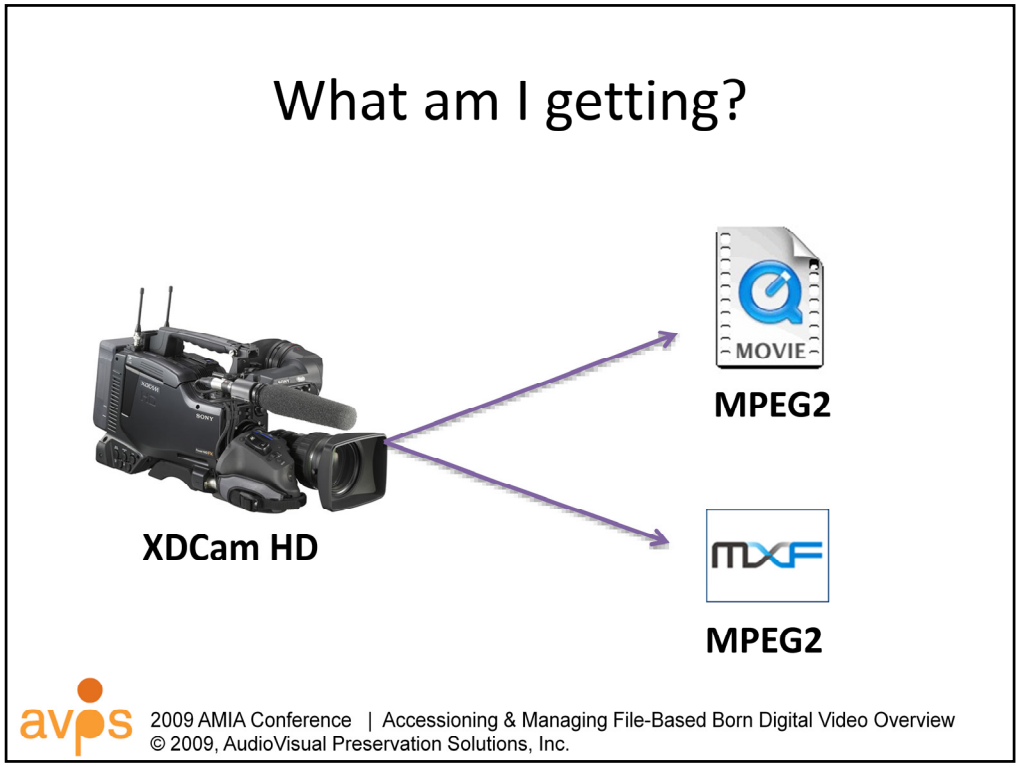
As examples of these …



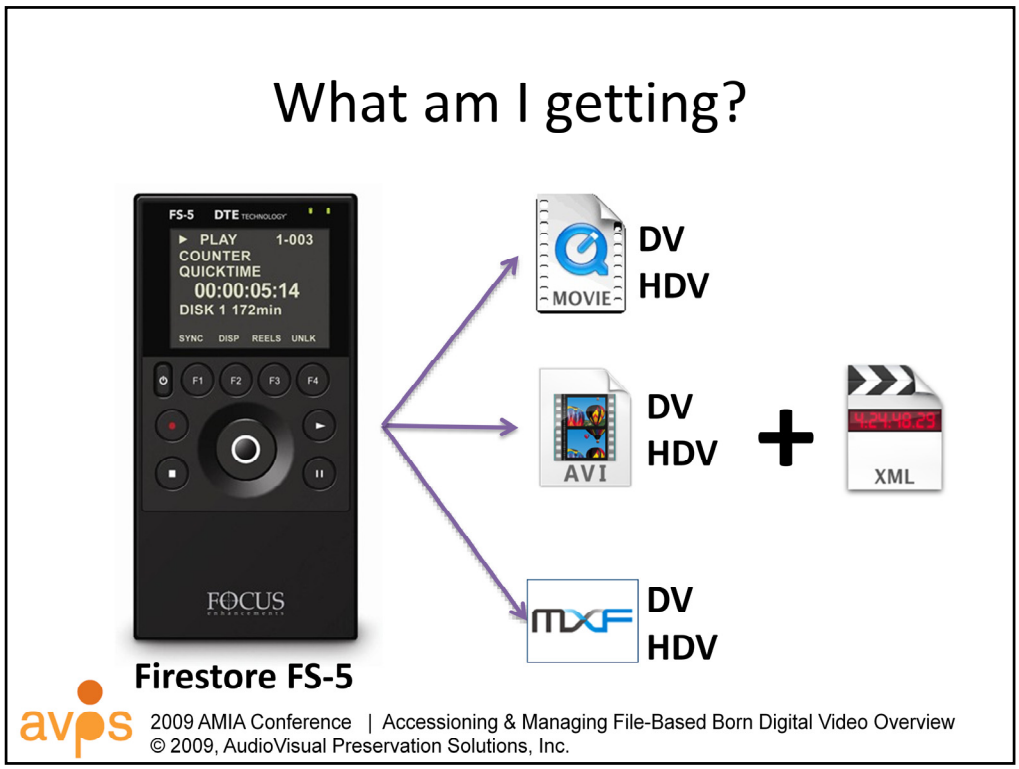
…we have everything from the iPhone and Flip cameras generating video these days en masse.



On the polar opposite end of things, the Red Camera is up to 4k resolution and will generate a proprietary R3D wrapper and codec, or you can generate a QuickTime wrapped proprietary codec.

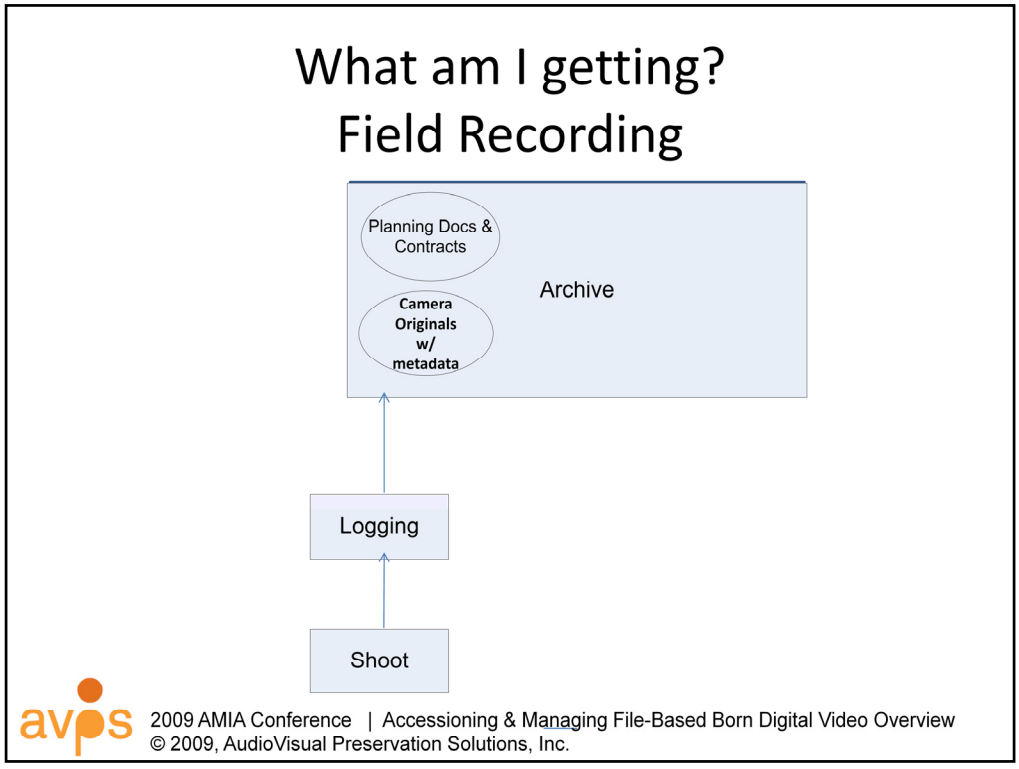


XDCAM HD generates Long GOP MPEG2 wrapped in MXF or QuickTime

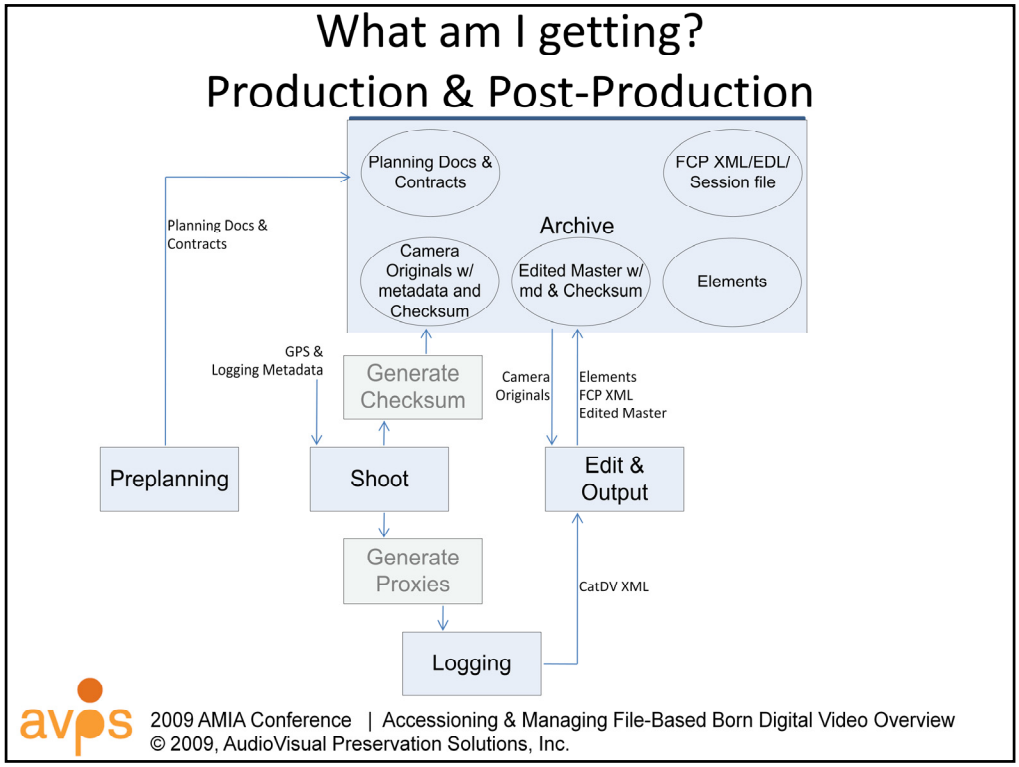


This is a device that receives a DV or HDV input, and allows wireless metadata entry during the recording process throughout a day of production which is stored as a sidecar XML file.

In thinking about the source as relative to the content being acquired it's worthwhile to give consideration to the following. Many files may see the end of the "production" chain after the initial recording. This would be typical for recordings of lectures and home movies. One would not expect external metadata documents to show up along with this type of content.



One step up in sophistication would be a field recording that is further annotated and logged after it is recorded. The formality of the scenario may also produce contracts such as releases.



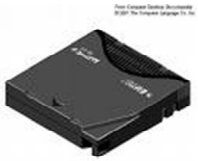
A more extensive production process creates deliverables including planning notes, production logs, transcripts, Edit Decision Lists and other metadata documents generated as a byproduct. Aside from the mix of video codecs and wrappers resulting from this process, one would likely see files being deposited in the form of text documents, XML documents, image formats, and an array of other application specific documents.

How am I getting it?

Hard Drive

FTP

Solid State Drive



Data Tape

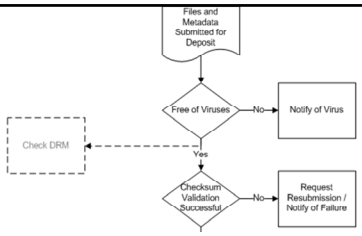
avps

2009 AMIA Conference | Accessioning & Managing File-Based Born Digital Video Overview

Typical delivery mechanisms for ***planned*** acquisitions include Hard Drives, encompassing the recent addition of solid state drives, data tape such as LTO, and FTP, though this gets less viable with large files.

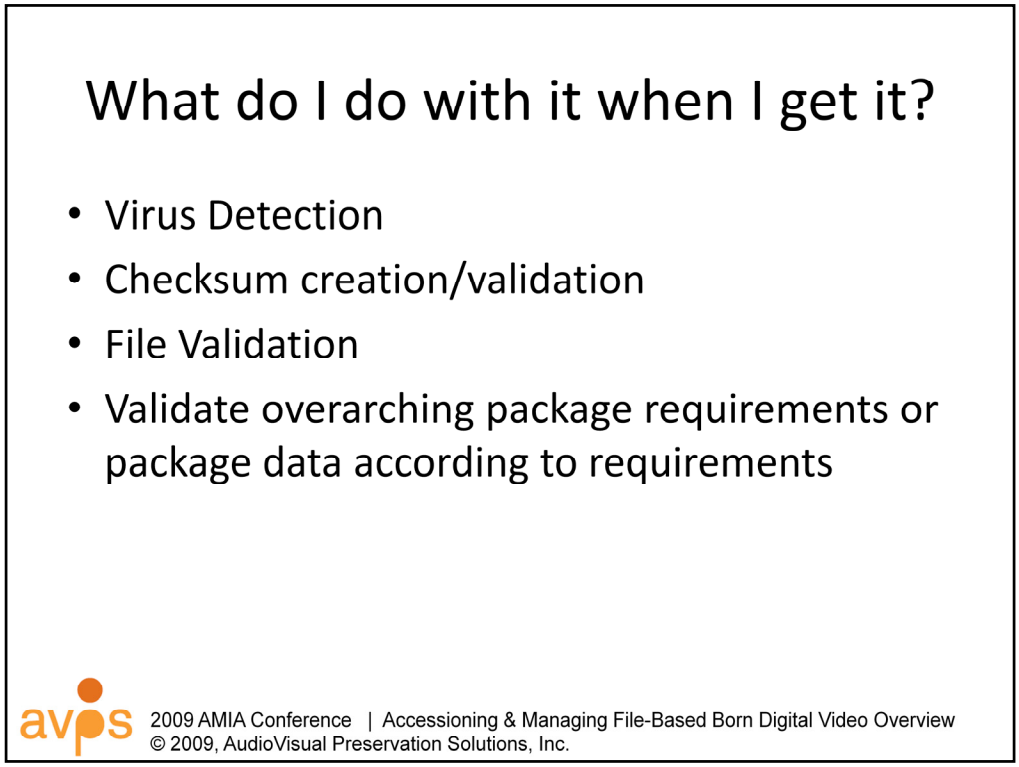
In thinking about acquiring content in lesser planned or unplanned scenarios there is overlap, but there's also a lot more! In scenarios such as acquiring a professionals "papers" after their retirement or acquiring the work of a video maker at the end of their career it is easy to imagine that born digital video will make its way into archives in the same way that all other personal digital documents do: On hard drives that have been sitting on shelves; on data tape serving as project backups with unknown necessary software interfaces from years past; on the devices themselves in the case of something like an iPhone or Flip camera; on CDs and DVDs; on proprietary media such as P2 cards or XDCam optical discs, and last but not least on multiple desktops and laptops that have been accumulated over the years and contains video content alongside all their other documents.

Hopefully you don't find any thumb drives that look like the one shown here!



WHAT AM I GETTING AND HOW AM I GETTING IT?

WHAT DO I DO WITH IT WHEN I GET IT?

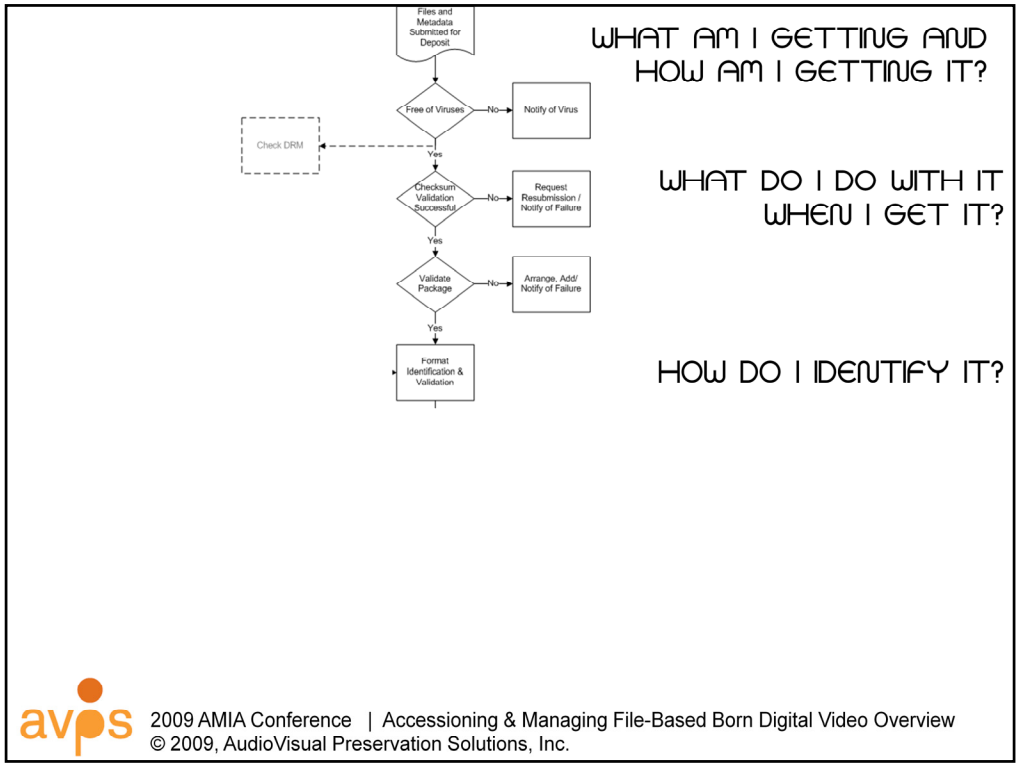


Given the convergence of video and everyday IT, viruses are a very real threat to archives accessioning file-based content. Making sure that you have virus scanning in place is very important.

A checksum for individual files (video or otherwise) or for a group of files may exist at the point of ingest. If present this will be in the form of a separate text file, usually carrying an extension incorporating the checksum type (e.g. xxxx.md5). The hash value also may be embedded in the file or in a database or spreadsheet deposited with the files. Running a checksum comparison will let you know if the data has changed (purposefully or otherwise) since the checksum was generated.

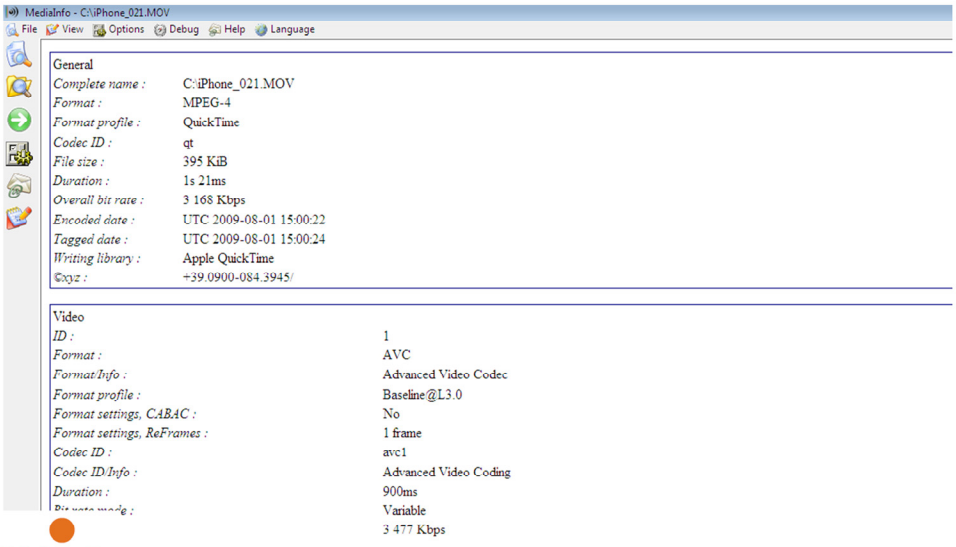
File validation ensures that the file is a valid instance of a document type according to specifications and standards.

Validating overarching package requirements implies that there are routines in place to unpack and process the package. In planned acquisitions, the archive may provide a specification for deliverables to the depositor. In unplanned acquisitions the content may arrive in total disorder and require organization and/or arrangement prior to performing any ingest routines.



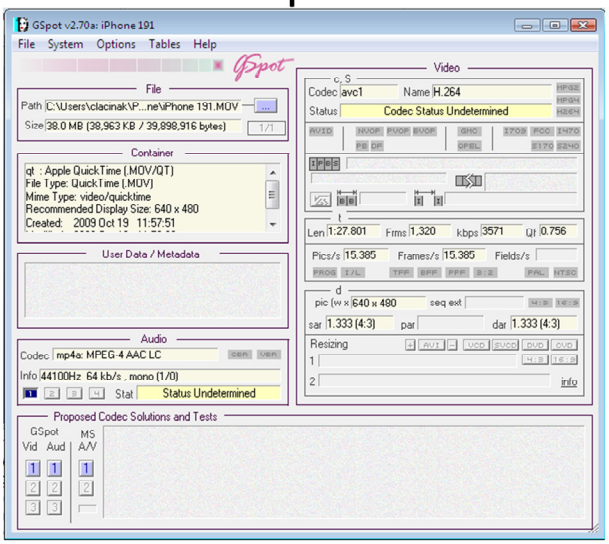
Identification in the file-based domain is imperative for the same reasons it is in the physical. It informs the management of the object and the toolset required to work with the file. Identification of media types, wrappers, codecs, resolution, bit rate settings, file size, file structure and embedded metadata is meaningful information.

How do I identify it?
MediaInfo



Identification can be achieved through the use of a wide array of tools, including MediaInfo. Notice the identification of embedded metadata and of wrappers and codecs. A note of interest is the GPS data listed for ©xyz.

How do I identify it?
GSpot



Gspot - reports codecs, wrappers and other technical metadata about a file. Video Inspector is similar.

How do I identify it?
ExifTool

ExifTool - Note the xmp embedded metadata reporting. All of these tools can be incorporated into ingest routines and workflows to report on the file properties, embedded metadata, attributes, structure and other technical information.

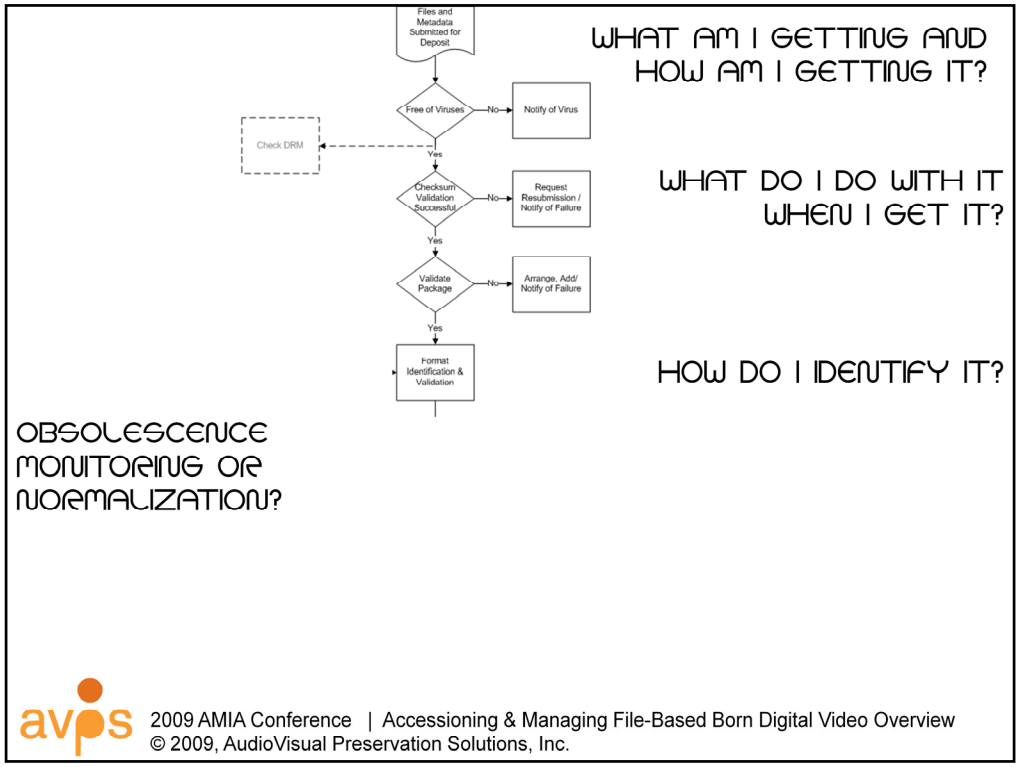
How do I identify it?
External Metadata – FCP XML

This information and other structural metadata may also be found in accompanying documents such as Final Cut Pro XML, and inferred through the directory structure. As opposed to embedded metadata, external metadata may identify how an object fits into a group of files.

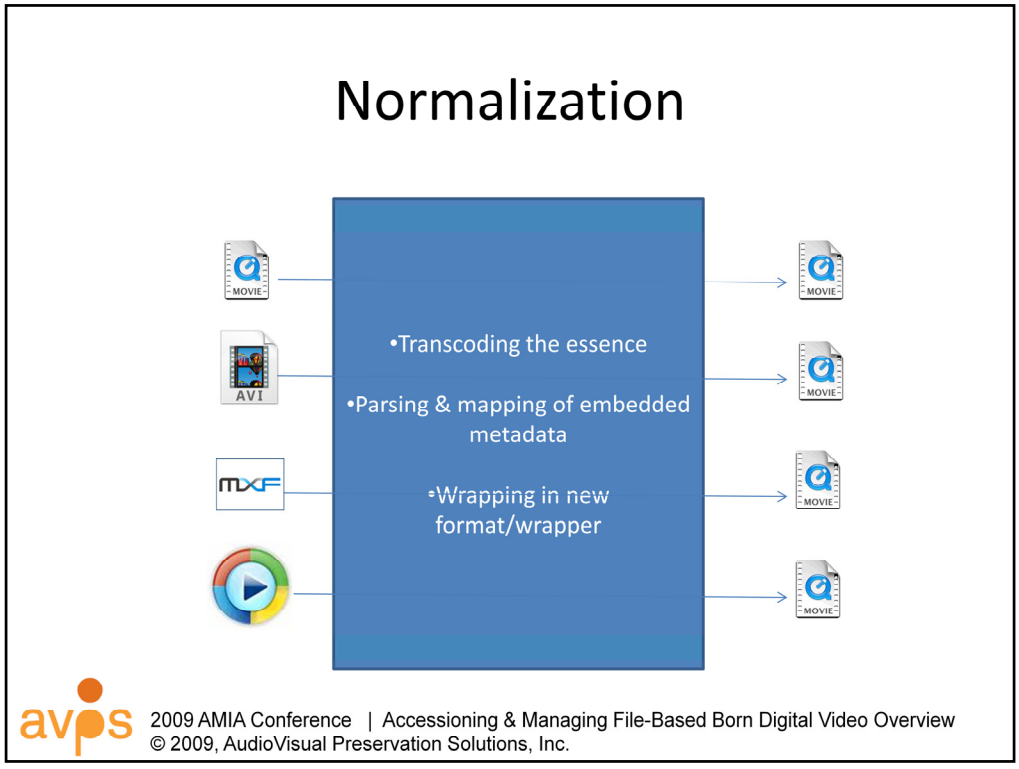
How do I identify it? Validation

- JHOVE

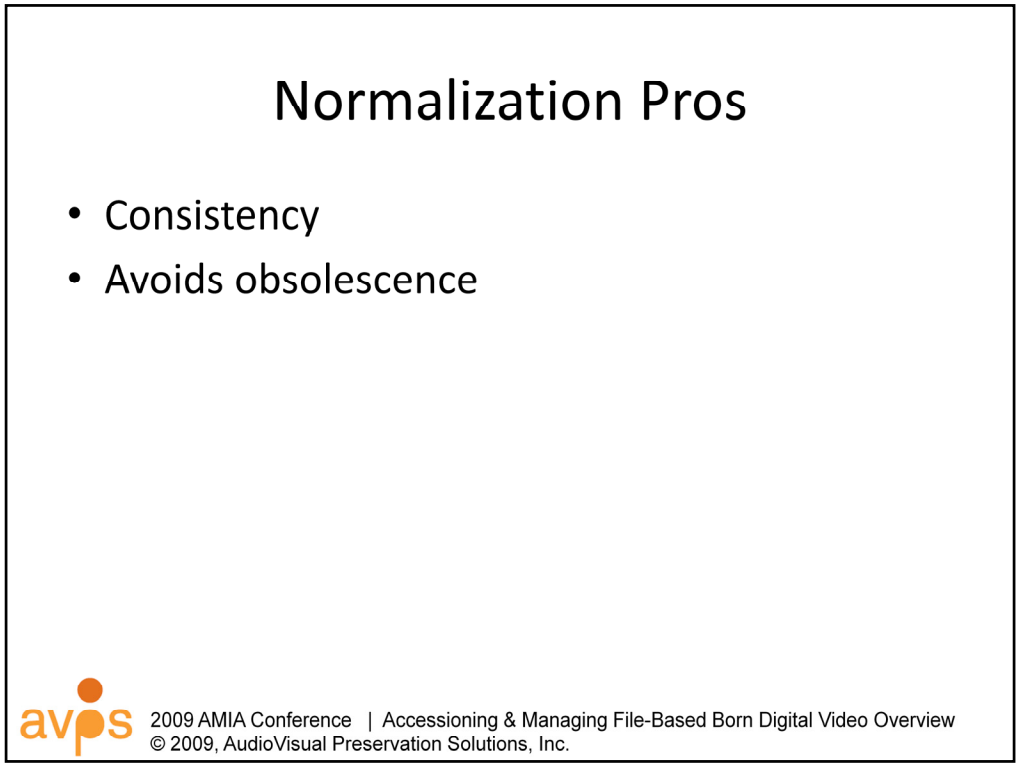
The fact that a file has a file extension labeling it as a certain file type doesn't mean it's a valid instance. Validation evaluates and reports on the conformance of the file with it's associated file type.



One of the reasons for identifying the file and it's properties is to assess the risk of obsolescence associated with that file. There are two approaches which are commonly discussed when it comes to accessioning born digital audiovisual content and managing the risk associated with obsolescence. These are Normalization and Obsolescence Monitoring. Let's take a look at these.

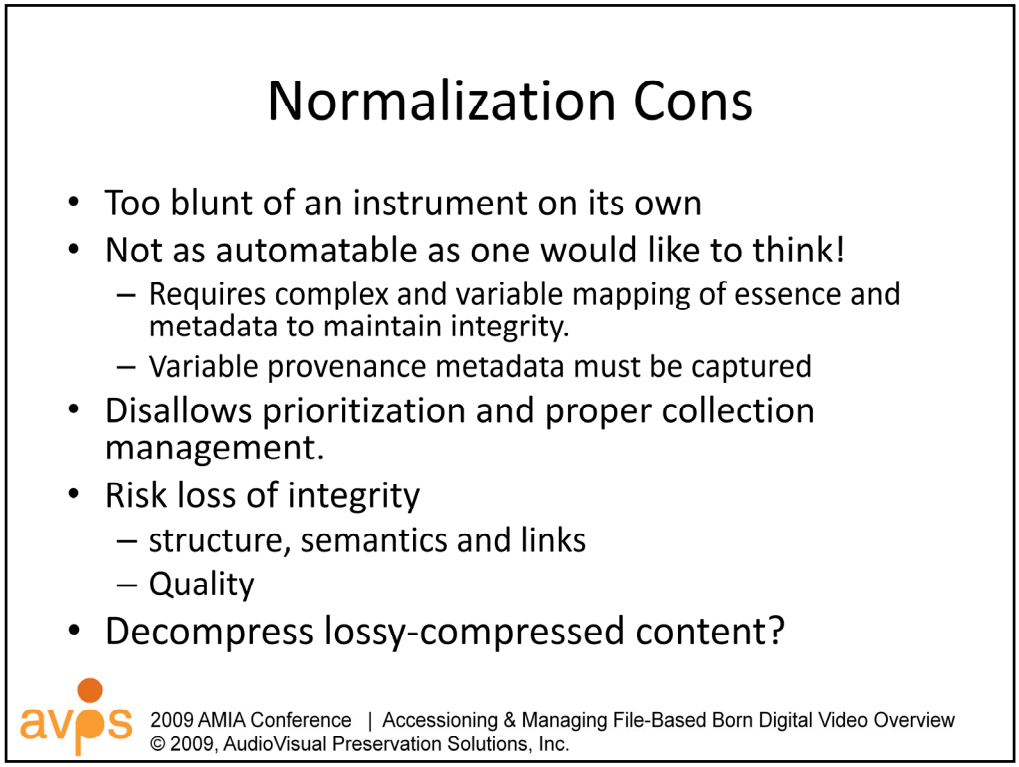


Normalization takes place at the point of ingest and is essentially proactive migration. At the point of ingest, files are transcoded to one consistent file format.



Consistency, is a great friend to preservation. It is good not only for immediate and near term collection management, but especially for the next migration which will take place. It is a nice ideal.

With proper selection of a target normalization format we greatly reduce the risk of obsolescence associated with other formats.

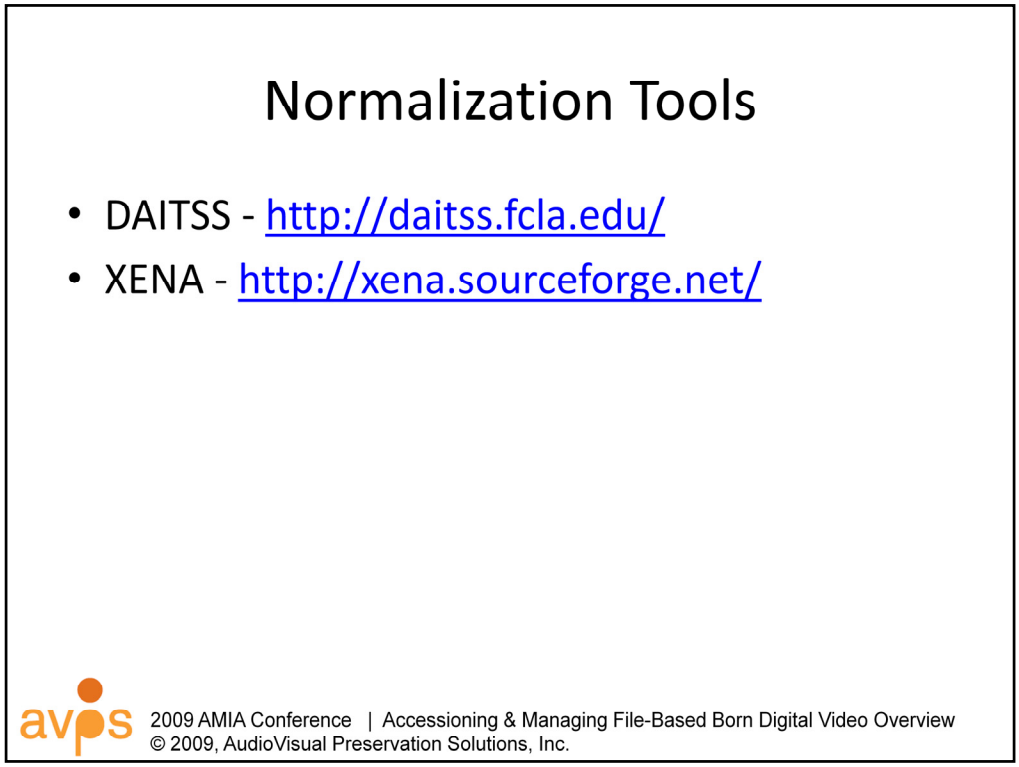


Ultimately, the cons are many. First, after scanning down the list of cons you will notice that as a sole tool, it's too blunt of an instrument. The reality is that with variable input of codecs, wrappers and metadata the task of creating the transcoding and mapping routines is no small task and it's fraught with difficult decision making and challenges. Especially as our digital packages, necessarily become more complex. The risk of lossy migration when done en masse like this is real.

On its own, it precedes attainment of intellectual control and repeats an error of the physical domain by expending resources in an imprecise manner.

Technically speaking we face even greater concerns. For anyone who has been involved in a mapping project you know what a challenge it can be to maintain structure, semantics and relationships. These might include breaking connections with siblings and peripheral dependent documents, losing markers, losing metadata structure and granularity, and more. Additionally, when tools are being used in a broad sweeping manner with variable input there is a real risk of poor and lossy transcoding and mapping, creating a low quality output as our master.

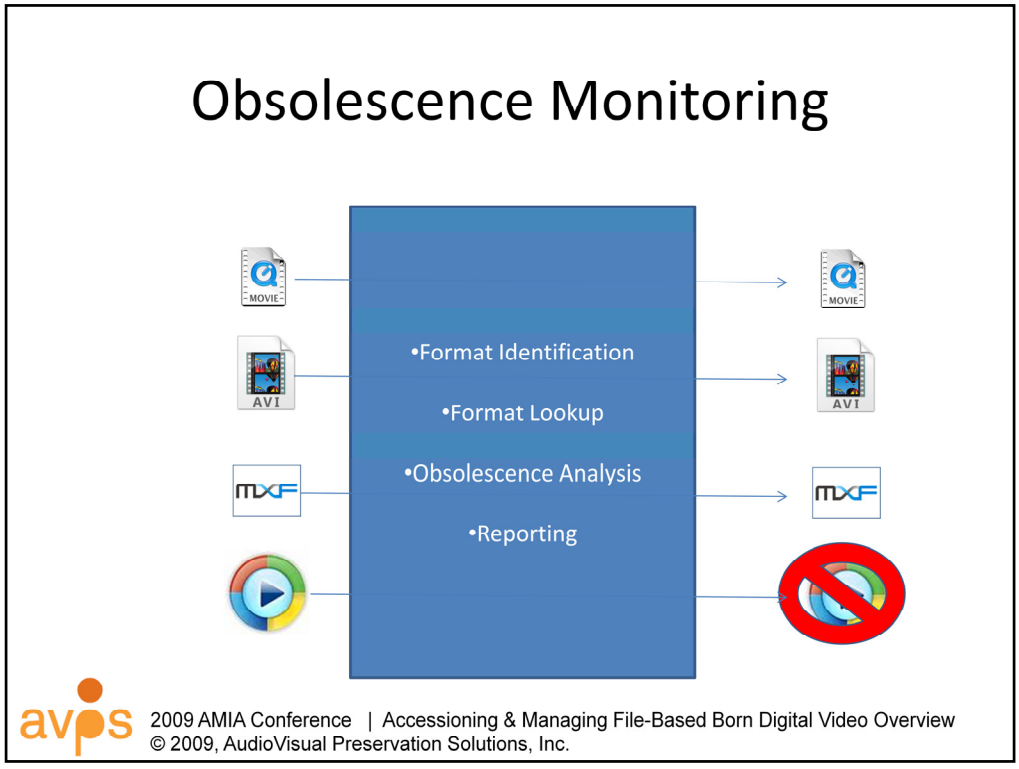
Lastly we see a practical consideration of storage. The likely candidate for a normalized output would be uncompressed, while many of the input formats would likely employ lossy compression. The result is the worst of both worlds. The negative of a larger file coupled with the negative of poorer audiovisual quality.



Here are a couple of tools that are available for normalization.

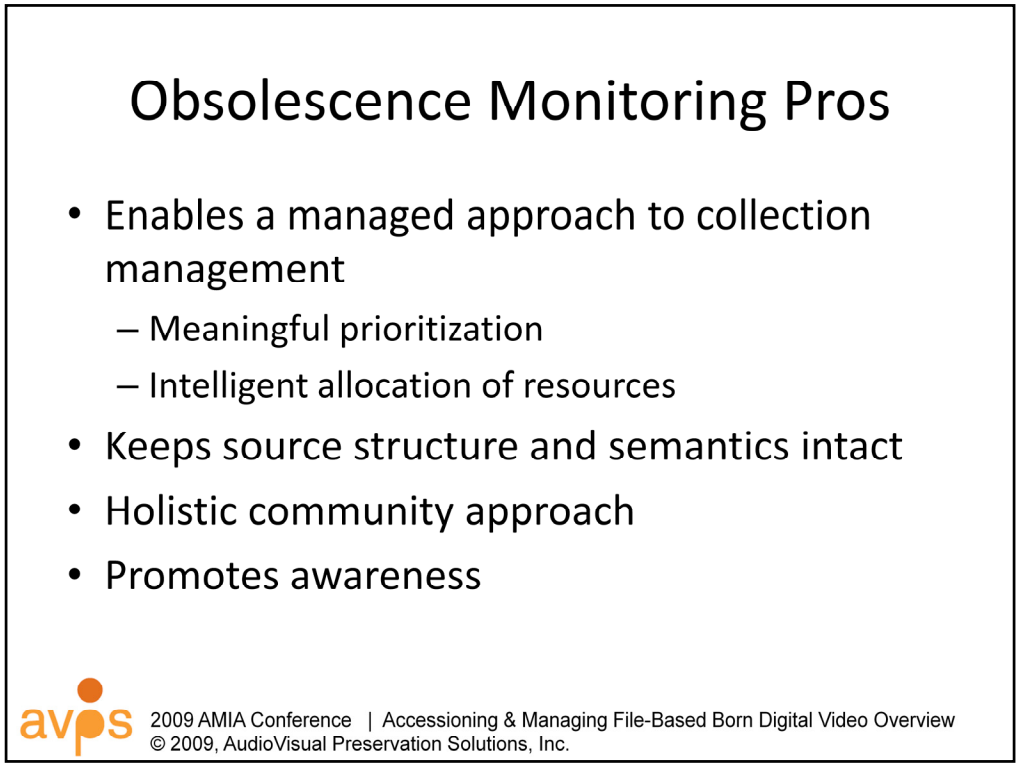
Tools that are used for normalization in larger organizations or by those with the appropriate IT skills are generally speaking single purpose command line utilities coordinated by scripts.

For smaller collections or organizations without access to the appropriate expertise, there are applications with GUIs (Graphical User Interfaces) that can be used in a batch mode to perform some of these activities, although you will hit an obstacle when looking for metadata mapping capability.



Obsolescence monitoring is a more sophisticated and complex mechanism, used at the point of ingest, at the point of access, in a retrospective batch mode and on an ongoing basis

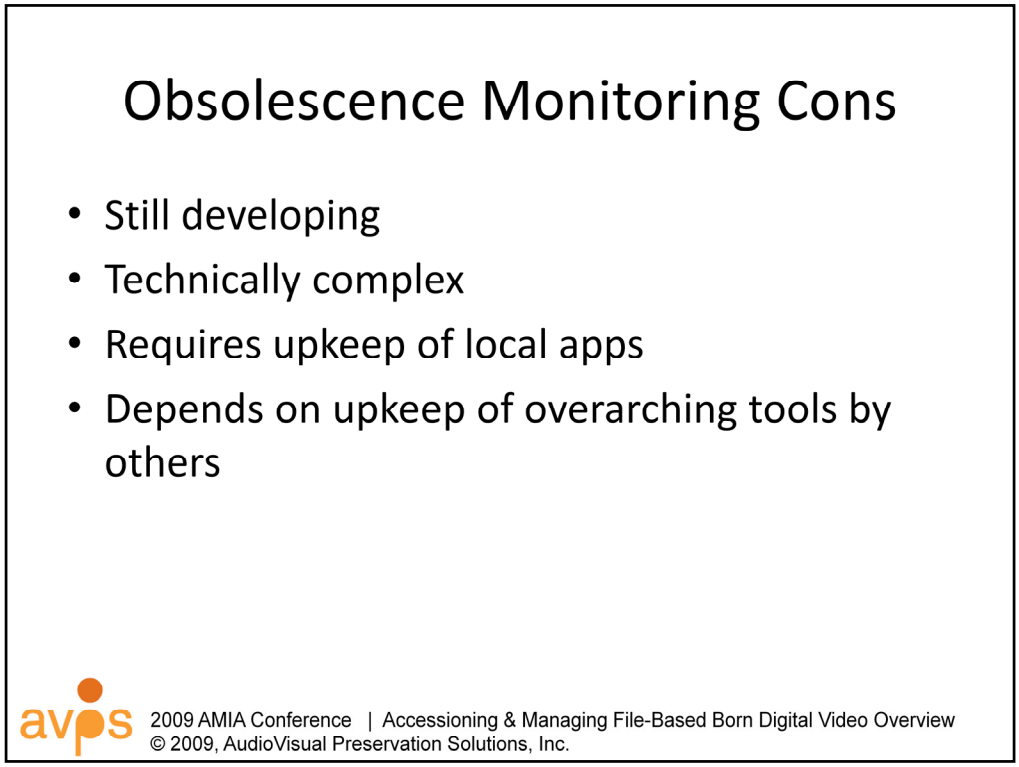
Simply stated, Obsolescence monitoring identifies and reports on formats that are at more immediate risk of obsolescence. However, the processes involved in this are more complex.



The pros of obsolescence monitoring are that it enables a much greater level of intellectual control where reasonable, while simultaneously alerting you to high risk issues. It places the prioritization and allocation of resources in your hands.

It also maintains the integrity of the original intact and allows for adequate investigation of the formats nuances and functionality prior to any migration.

It relies on the community to maintain these tools, which is noted as a risk on the following slide of cons, but the benefit is also notable. The fact that it takes a community of people also means that those people are actively engaging and necessarily tuned in to the issues that effect us all. It raises awareness and can provide impetus for ongoing community solutions.



The main con is that these tools are still very much in development. The integration of the tools and their exchange of data is in need of refining in order to function more smoothly.

It is also technically complex and requires some level of expertise to install and maintain the applications.

And as stated previously, this system heavily depends on others which is both a good and a bad thing.

Obsolescence Monitoring Tools

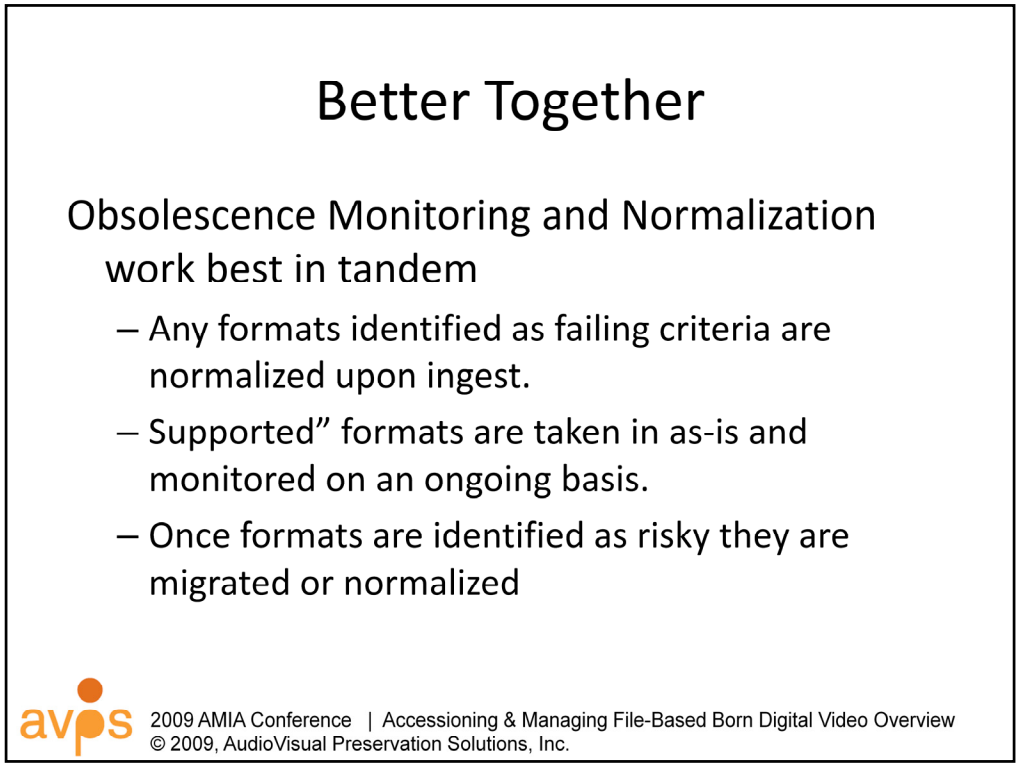
- AONS II (Automatic Obsolescence Notification System)
  - Download: http://sourceforge.net/projects/aons/
  - Blog: http://aons2dev.blogspot.com/
- UDFR (Unified Digital Formats Registry) - http://www.gdfr.info/udfr.html
  - PRONOM - http://www.nationalarchives.gov.uk/PRONOM/Default.aspx#
  - GDFR (Global Digital Format Registry) - http://www.gdfr.info/
- Library of Congress Sustainability of Digital Formats http://www.digitalpreservation.gov/formats/intro/intro.shtml
- JHOVE (JSTOR/Harvard Object Validation Environment) - http://hul.harvard.edu/jhove/
- DROID (Digital Record Object Identification) - http://droid.sourceforge.net/wiki/index.php/Introduction

Obsolescence Monitoring Small Scale

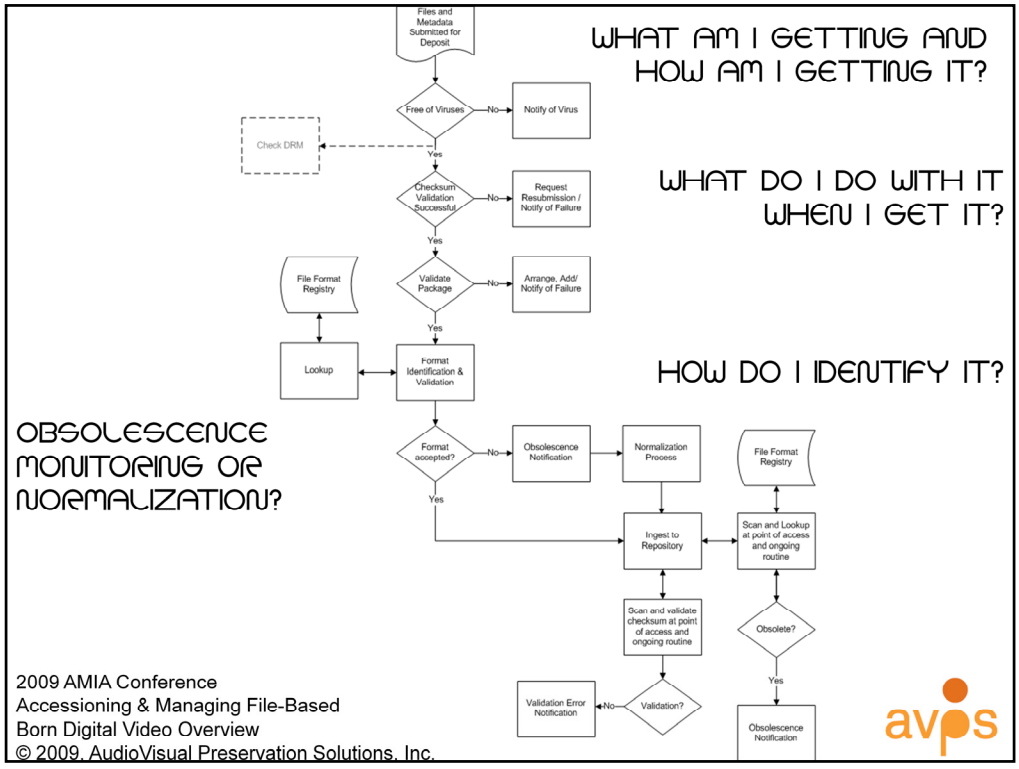
Common Sense Approach: Manually

- Keep plugged in to UDFR
- Apply Sustainability Factors and AONS criteria

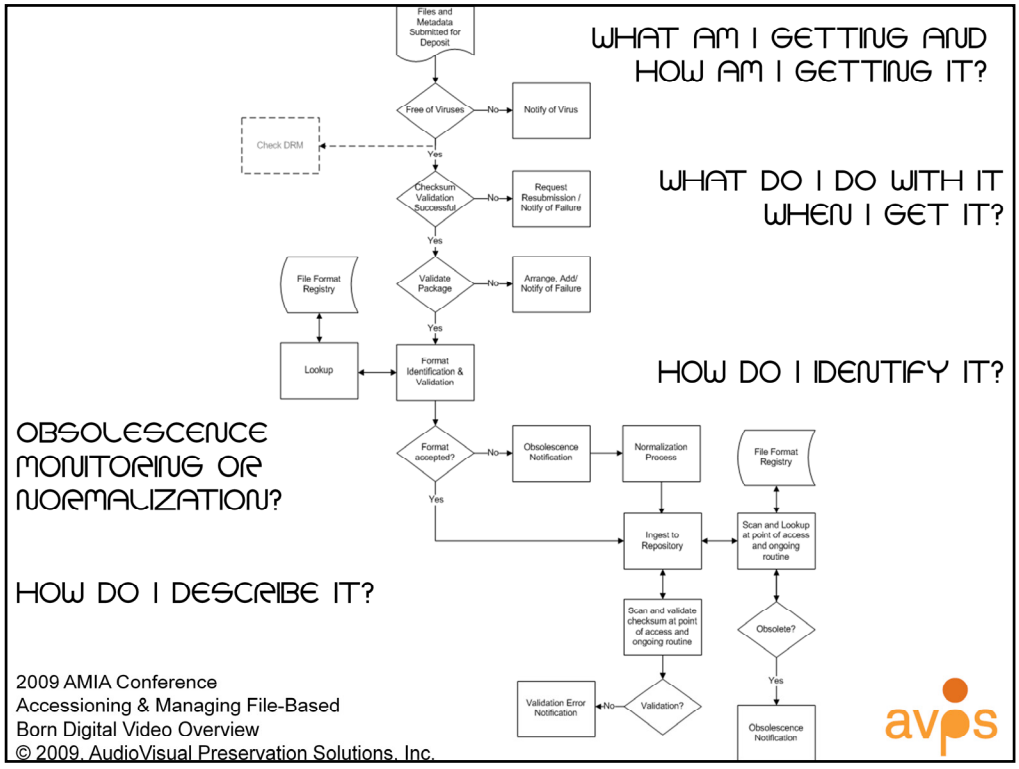
Recognizing the complicating factors just stated, obsolescence monitoring can also be performed on a small scale, or in situations where there is not too great a variety of inputs using some of the same tools manually. Less than ideal, but better than nothing.



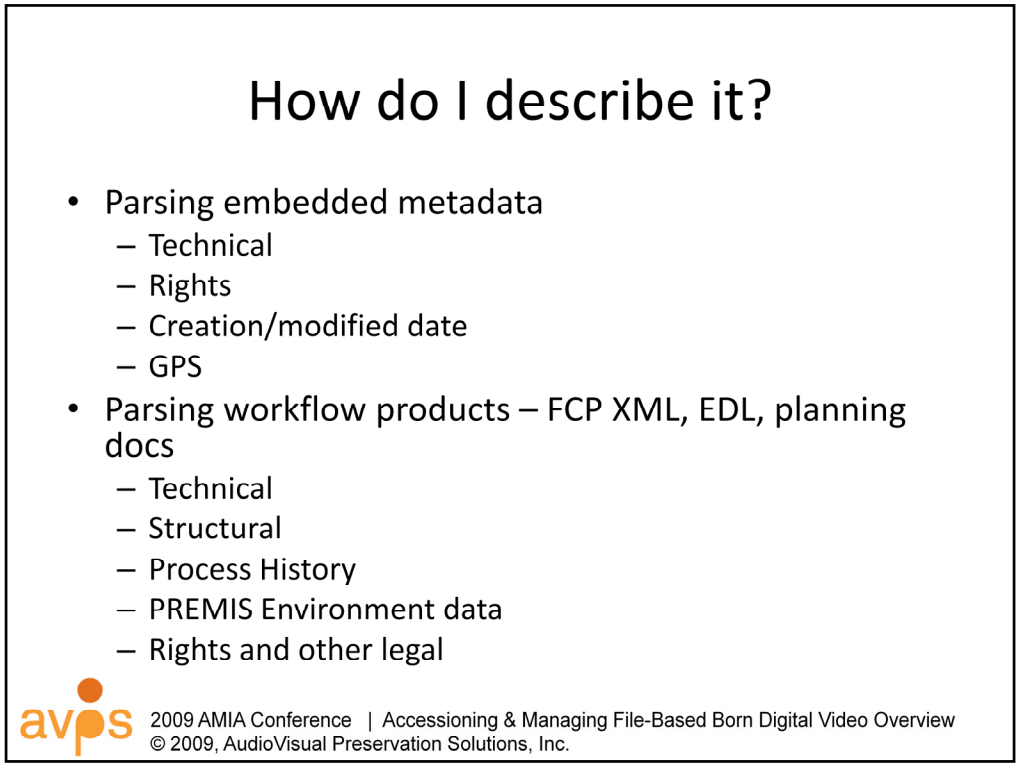
Normalization is too blunt of a tool to use solely and obsolescence monitoring eventually gets you to the question with high risk formats of "now what?", with the obvious answer of migration.



And showing the rest of the workflow shows the combined use of Obsolescence monitoring and migration



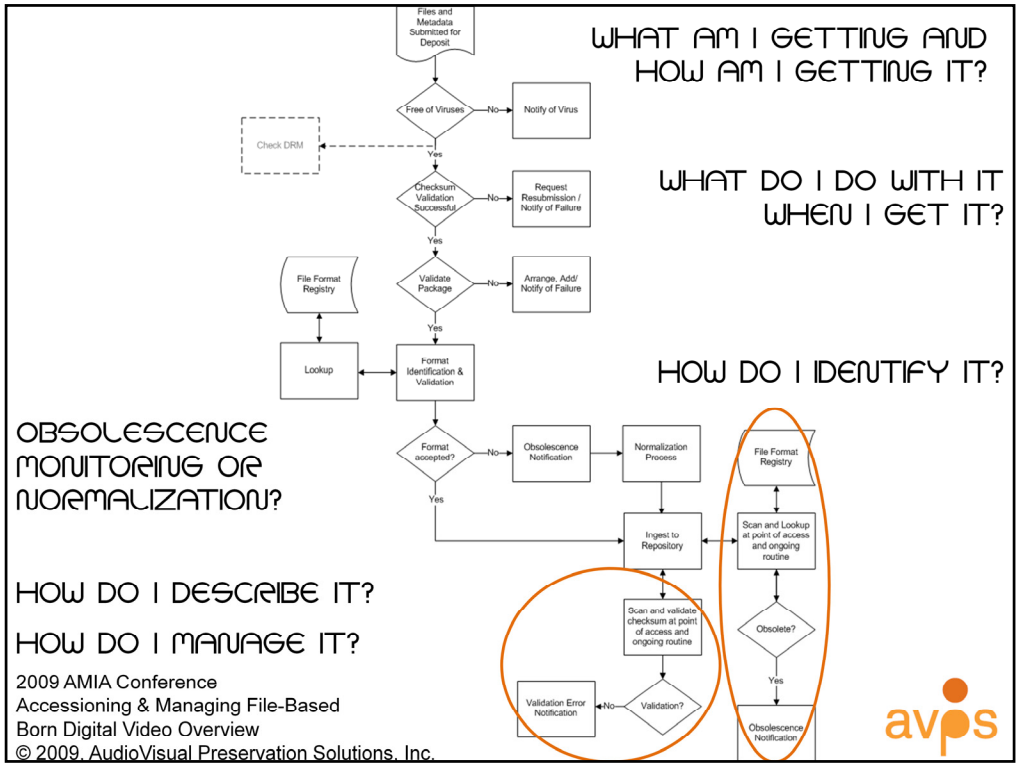
2009 AMIA Conference
Accessioning & Managing File-Based Born Digital Video Overview



Leveraging existing metadata in the deliverables is essential in efficiently and accurately describing the technical, administrative and descriptive characteristics of the file and its content. This gives us information about the object itself, it's relationships to other objects and provides other information critical to access, presentation and proper management.

Speaking of legal and embedded metadata

This is a recent case where the judge ruled that embedded metadata within documents was able to be used as evidence in court. This is an important example of the growing awareness and significance of embedded metadata, its persistence and integrity.



Assuming that best practices are followed for storage, including redundancy, geographical separation and refreshing there are still considerations to take into account. Active management is a necessary part of preservation and access. Among necessary management activities are:

•Ongoing obsolescence monitoring and planned migration at pertinent points.

•Ongoing automated checksum validation as well as validation at points of retrieval and access.

•And I don't have it listed here, but updating of metadata according to planned routines and at points of retrieval and access. This is necessary to maintain a reconciliation between metadata stored in deep storage and that which is stored in readily accessible and editable databases and other documents.

Feel free to contact me at chris@avpreserve.com and visit our website www.avpreserve.com for more information.