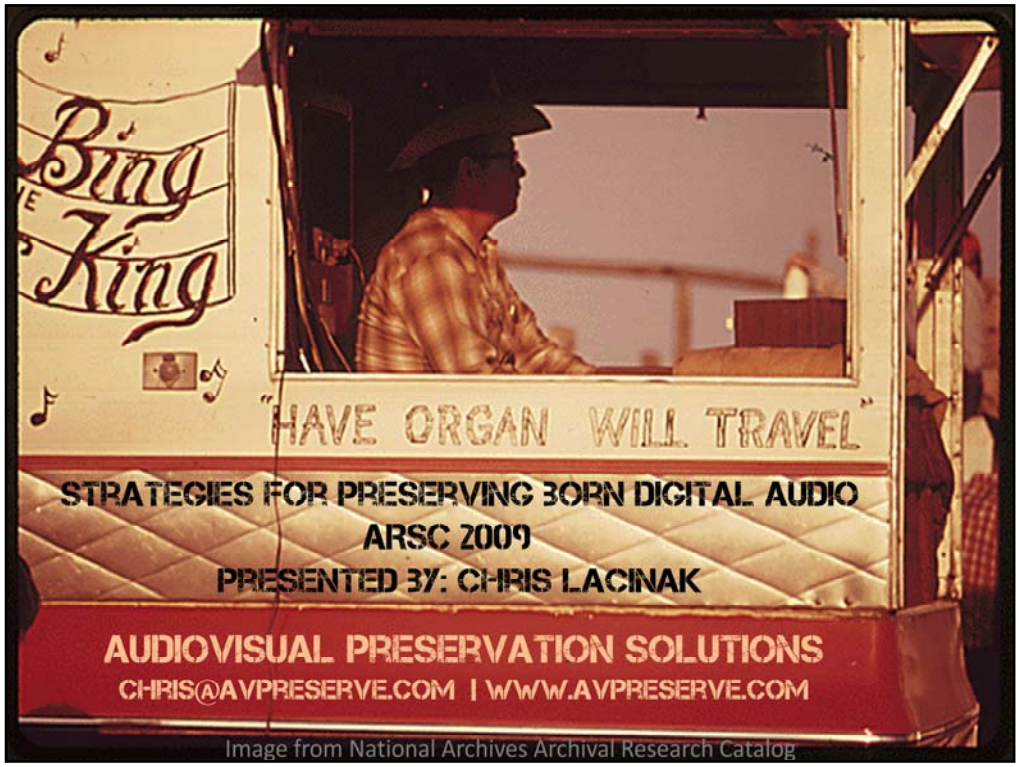


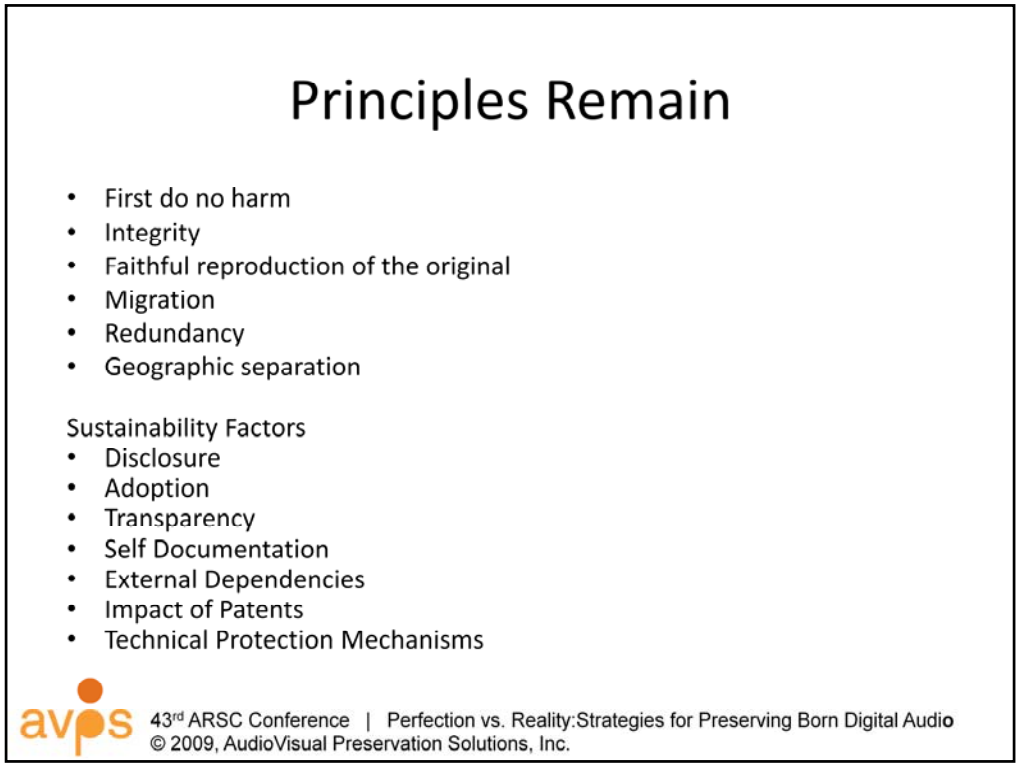
I should clarify up front that while the overarching topic is Digital Preservation Strategies, there will be a focus on obsolescence monitoring and normalization. The topic of obsolescence monitoring and normalization came to be singled out as a relevant topic to discuss through discussion within the ARSC TC a few months back. There are obviously so many great topics to discuss that fall under the umbrella of preservation strategies. Given the time I will only touch on a few others aside from the aforementioned.

A couple of notable things I am not talking about, but deserve special mention and awareness when thinking of born digital:

1. content living on data tape which uses 3rd party software as the access point to the data. This software becomes obsolete rapidly and if recovery is able to be performed with independent utilities it is very labor intensive.

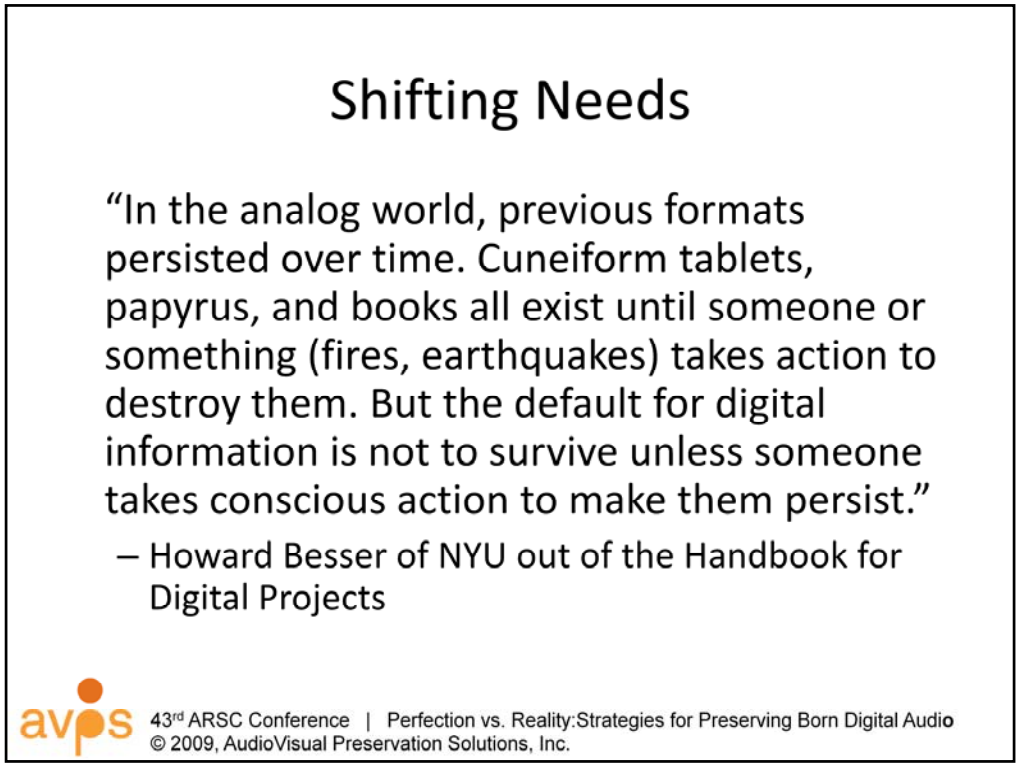
2. The backlog of ProTools sessions and the like, which are at great risk of loss unless they are dealt with immediately. Moving forward we have some great recommendations from NARAS and AES on how to deal with this, but there is a large quantity of at-risk content prior to the existence of these recommendations.

These are topics for a different presentation. I bring these up now only to put them on your radar if they are not already.

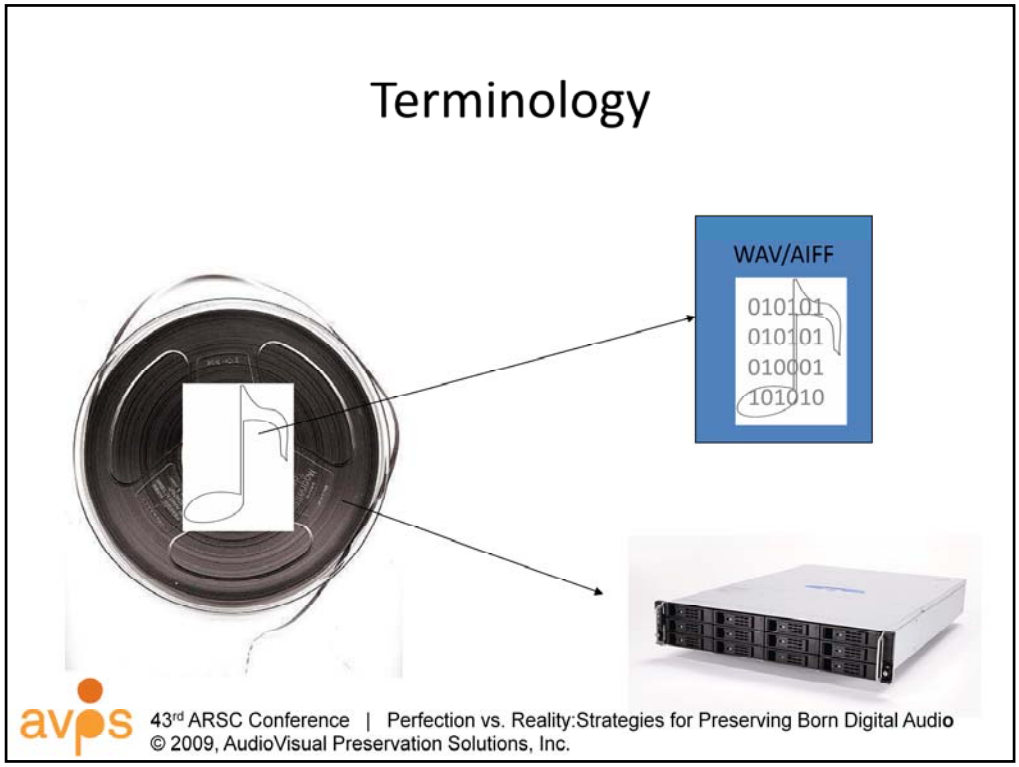


When preserving digital file-based content, we must adapt our focus and thinking accordingly. It requires the consideration of a different set of parameters and activities apart from the preservation of physical assets.

While many things change, principles remain the same, such as those listed here. These are principles that were true 50 years ago and will still be true in another 50 years, but how we apply them will undoubtedly change.

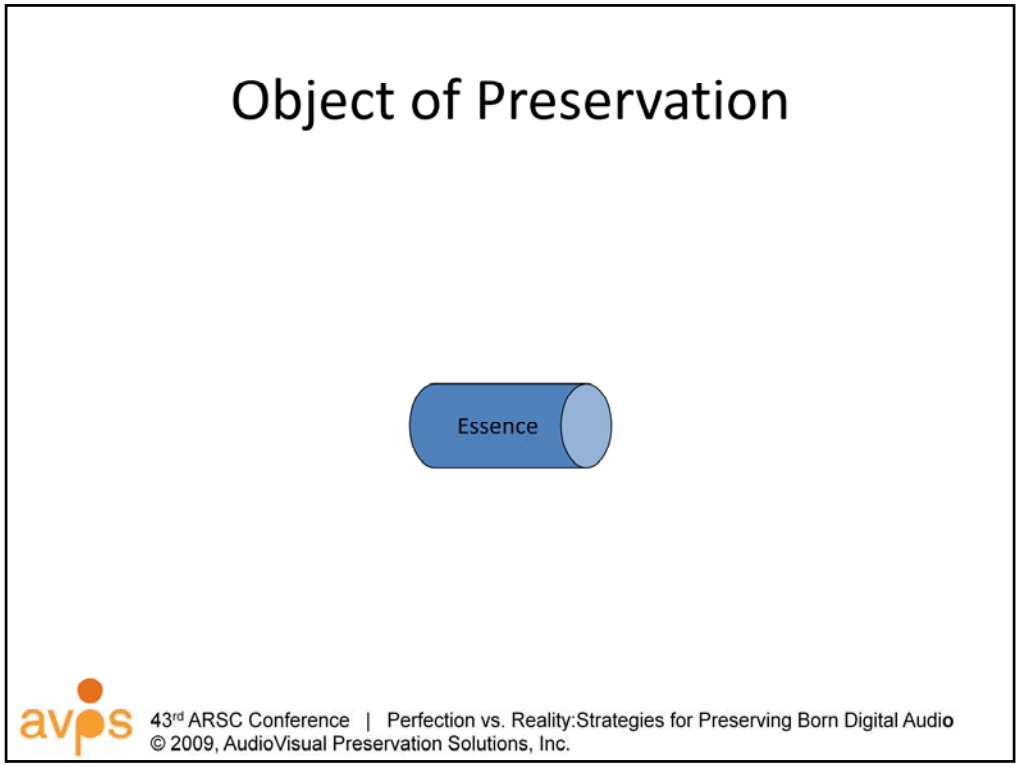


And needs change. This quote from Howard Besser sums up one of the primary shifts in need. We will not be allowed the same latitude in the file-based domain as we were in physical analog domain. File based content demands active management to persist. Fortunately along with this new burden, file based content brings along with it many opportunities for automation and efficiency that we never had in the physical analog domain. Aside from recognition of *this* shift, interpretation of principles into Practices, skill-sets and terminology also change.

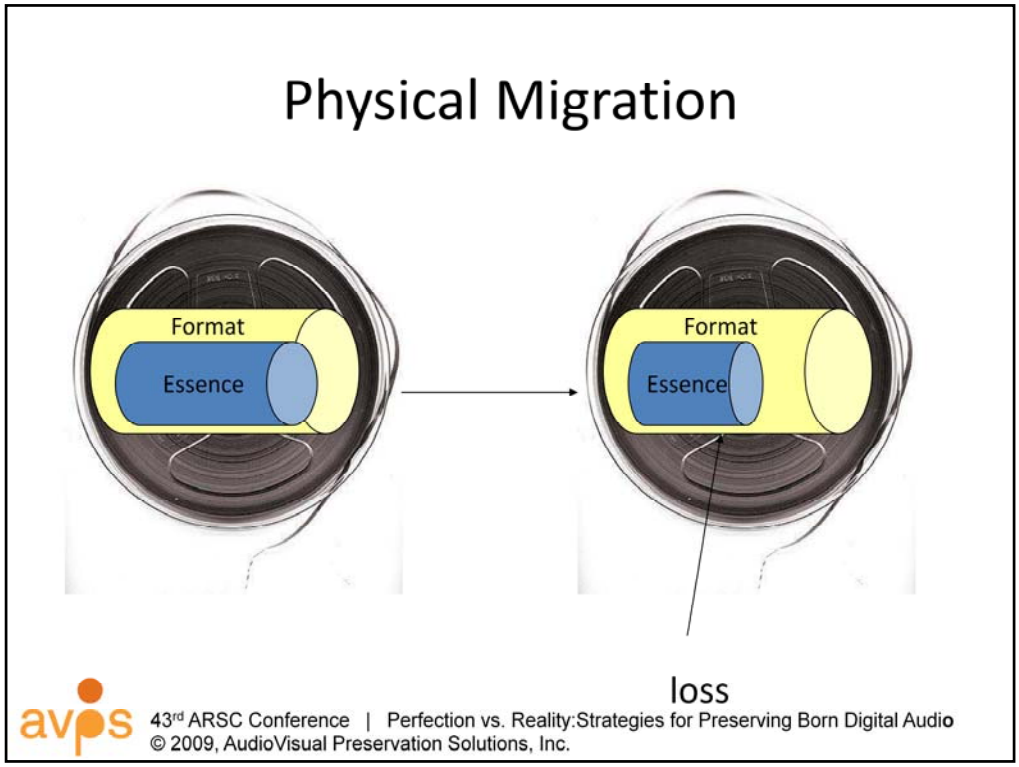


As an example of a shift in terminology with the focus on the principle of migration combined with the arrival of file-based content, the distinction between the intangible object from the tangible object became meaningful (the difference between the sound that constitutes a track in a recording and its separation from its carrier.)

Preserving a recording once meant solely preserving the carrier, and while there is a strong and direct causal relationship between the carrier and the essence, stated simply, we have realized that the object of preservation is indeed the essence. That the persistence of the essence is the goal.

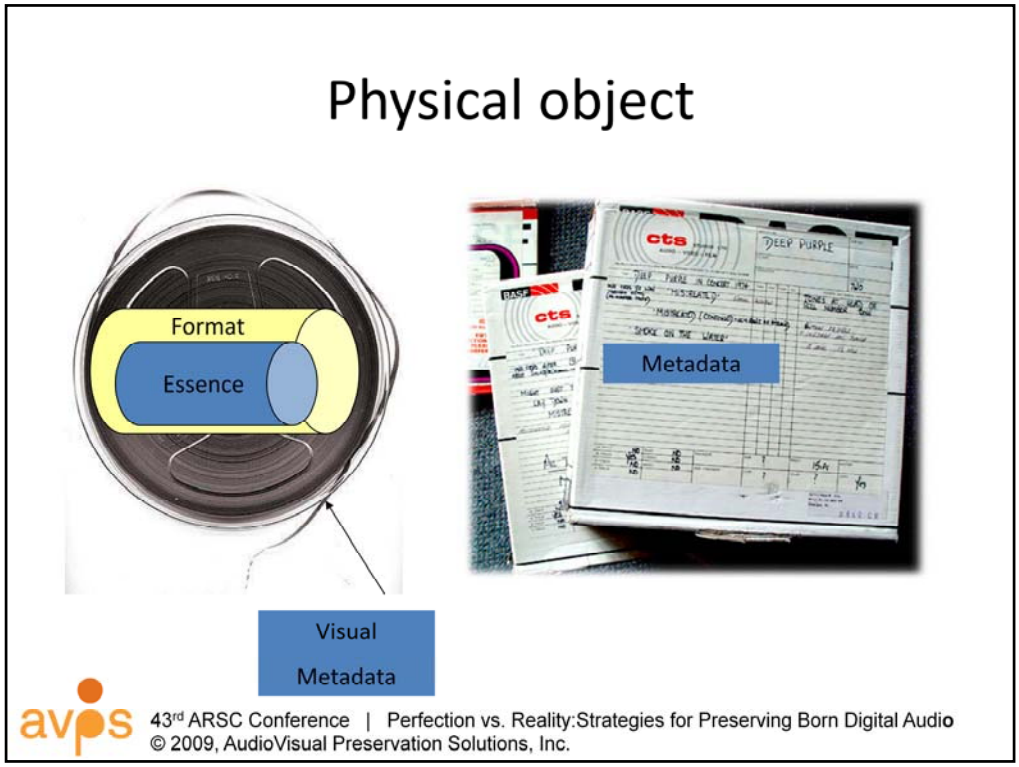


So here is our essence, separated from all media and formats. Separation of the essence from the carrier causes a paradigm shift in the actions taken to preserve a recording. The question is no longer “how can we ensure the longevity of a piece of media”. It is “how can ensure the longevity of the recording”. The answer to each of those questions is very different. They both involve migration, but approached from different perspectives. Migration was once an act of last resort. In the digital domain it is a fundamental routine in the preservation of content. In a properly functioning digital preservation system the media object is of little value. Assuming we have the proper infrastructure in place we do not depend on an individual hard disk. A hard disk should place no signature on the content and the survival of the content is not dependent on one hard disk.

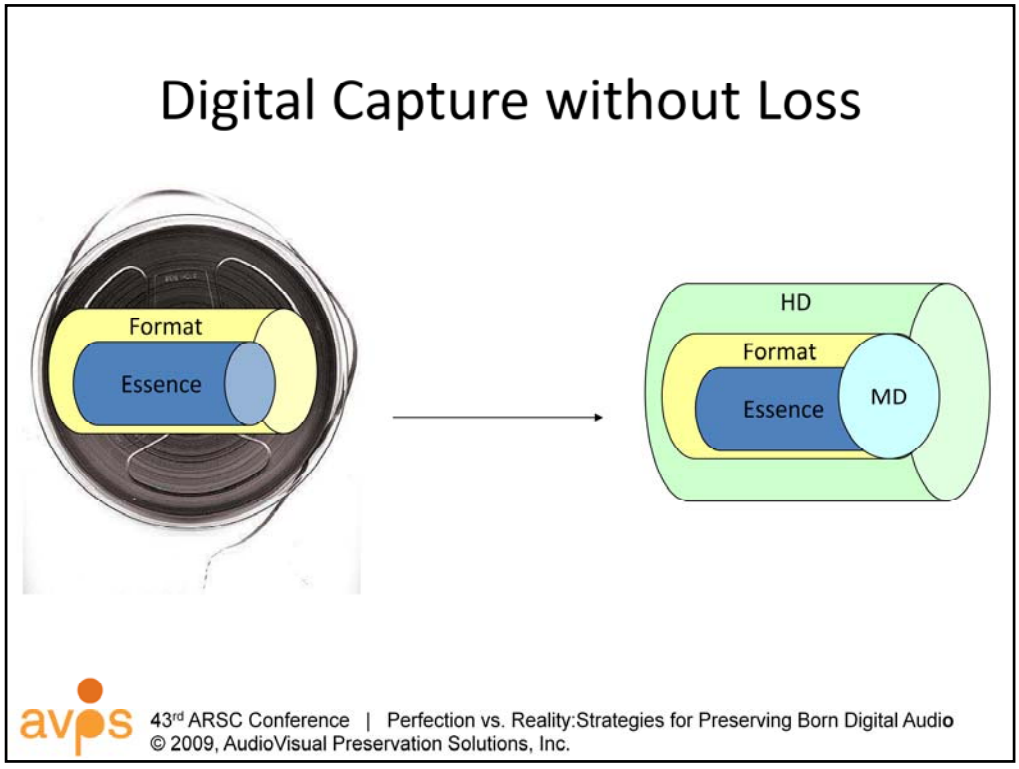


When we worked only with physical analog and early digital technology the recognition of separation between essence and carrier was not useful.

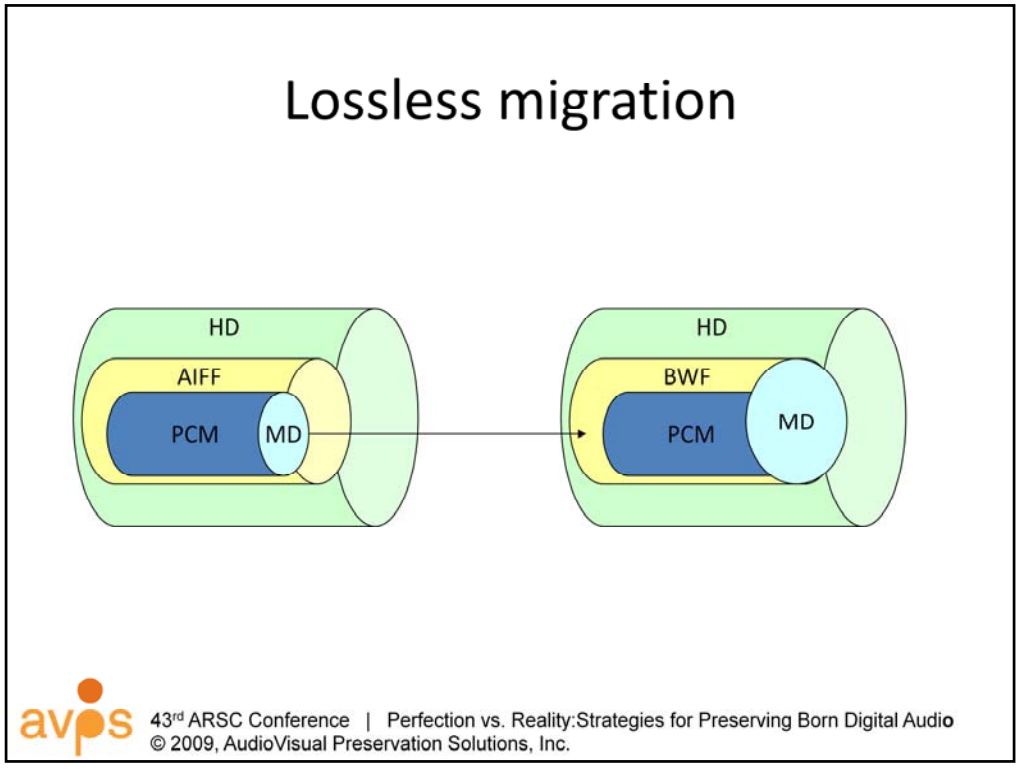
This was due to the fact that the processes used to migrate the essence were inherently lossy, which goes against the preservation principle of integrity and interest in quality.



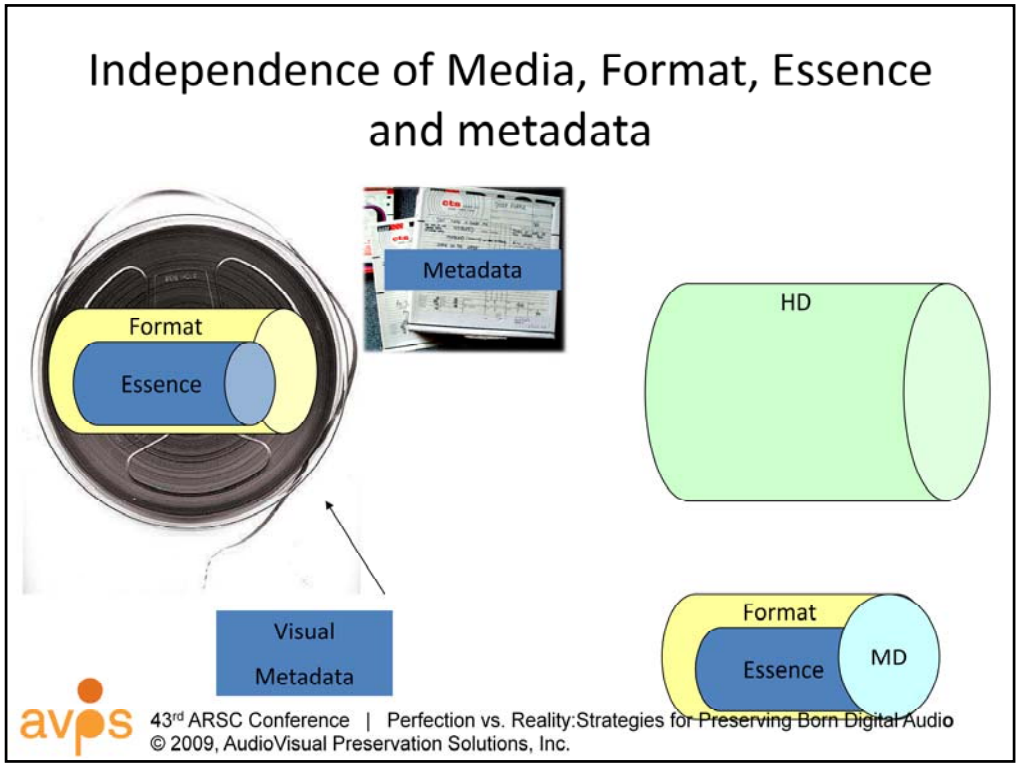
This is where we find the logic for reliance on the media. All actions taken and choices made in an effort to preserve revolved around the longevity of that piece of media. When a choice was made about media an imprint was being made on the resulting sound recording. From that point on reformatting would mean loss, so all focus revolved around the media object and the technology that revolved around it.



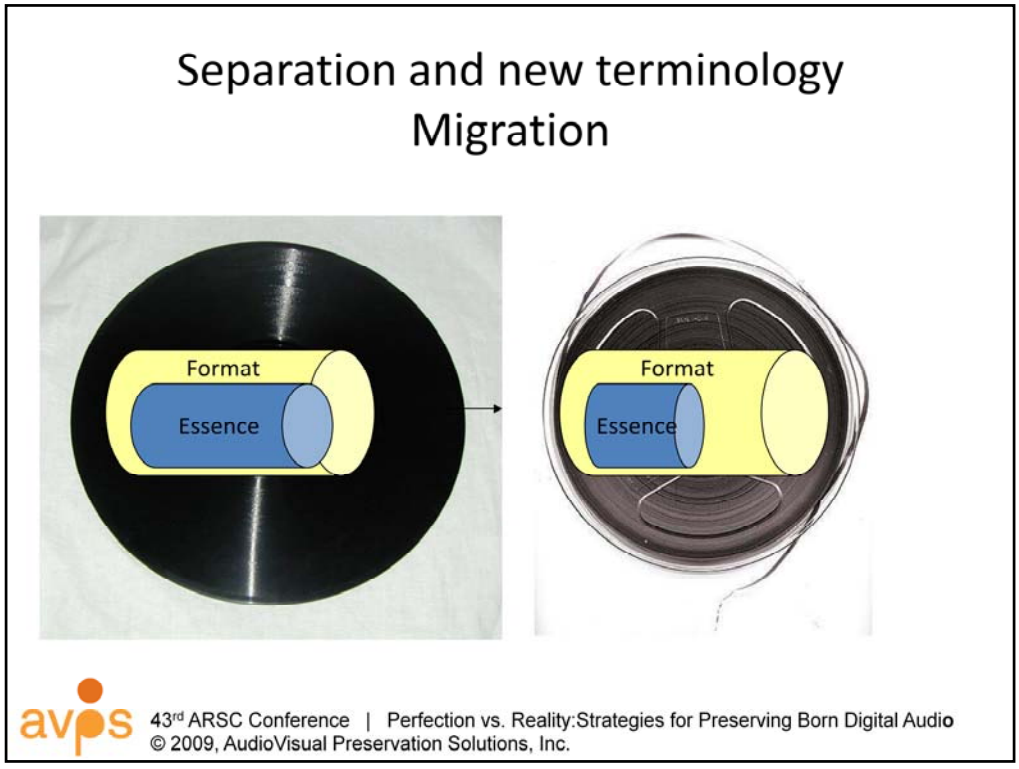
It was only once digital technology evolved to the point that it could capture an analog object without loss



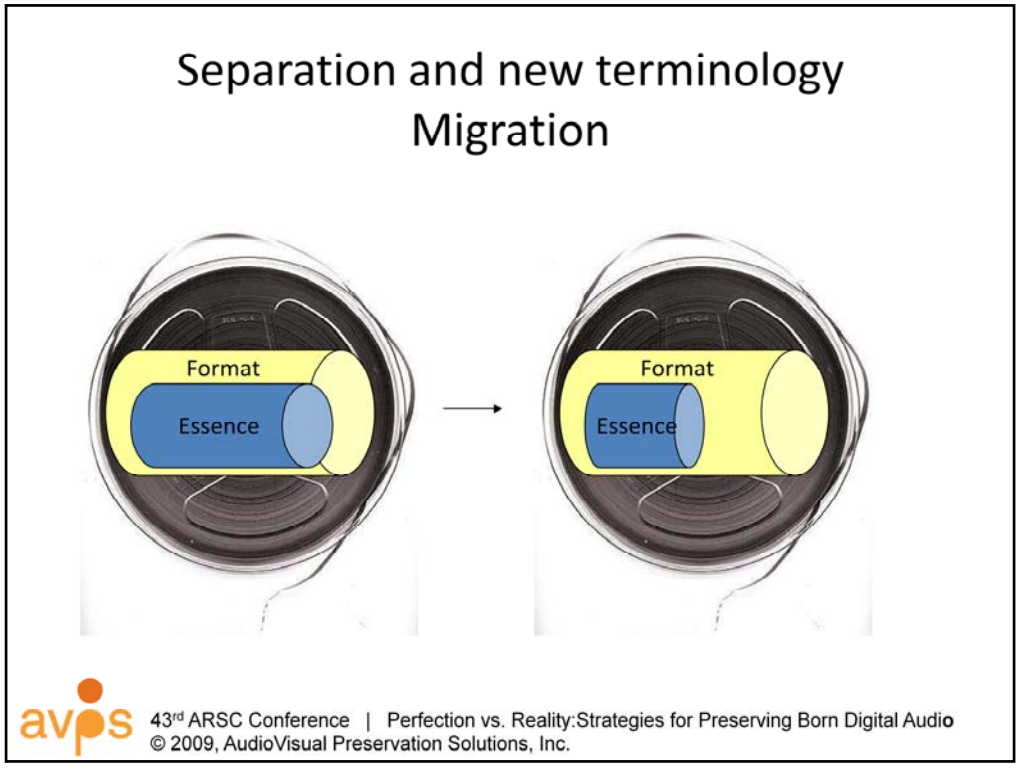
and also migrate those bits without loss



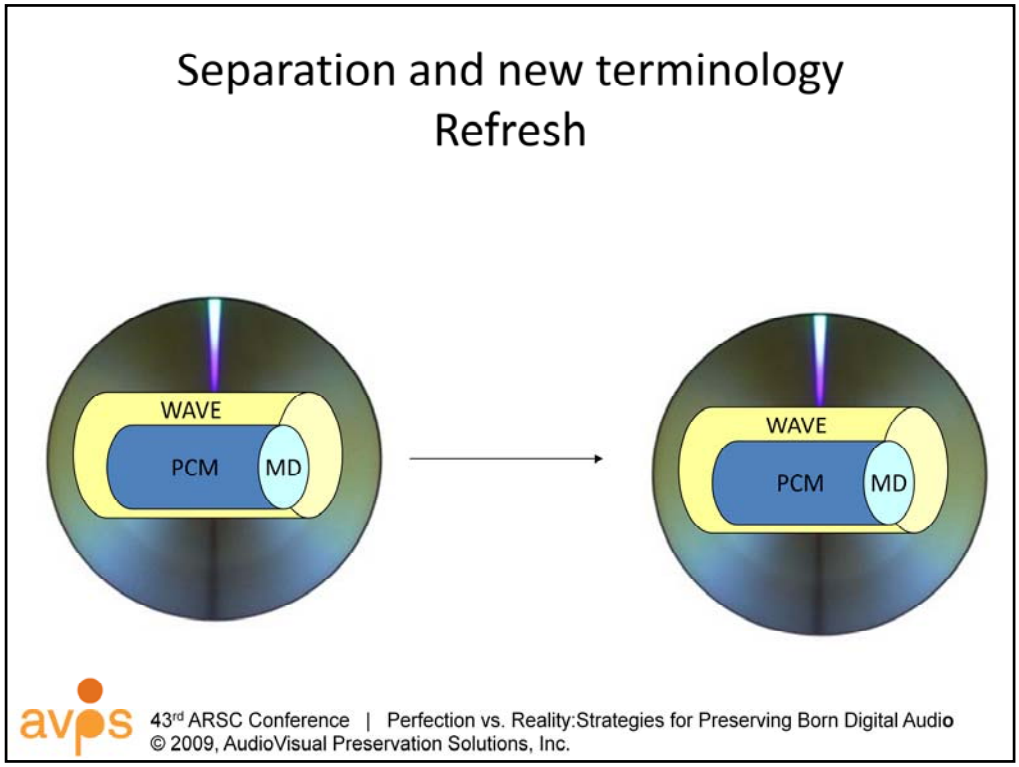
that the separation between essence and the object became a meaningful observation.



Along with this understanding came a division of terminology where migration turned into two new terms. migration and refreshing. There was no distinction previously. Both of these examples were called migration and constituted both refreshing and migration. From disc to tape

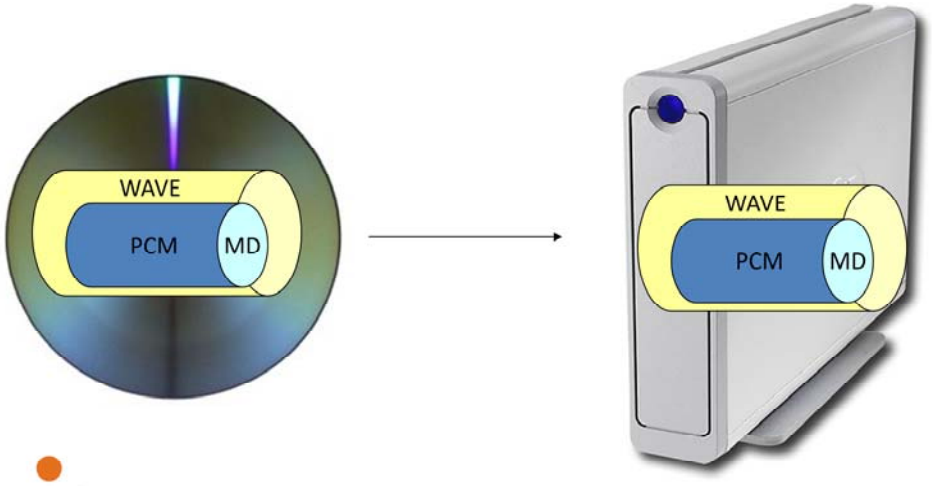


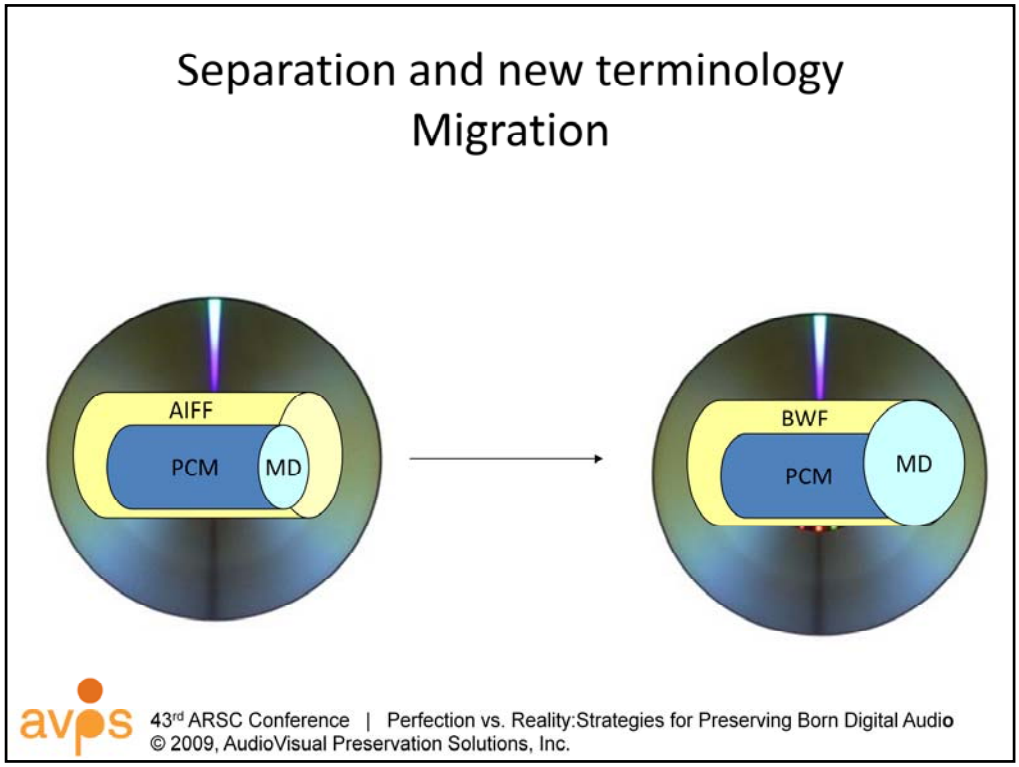
And from tape to tape were both considered the same essential process.



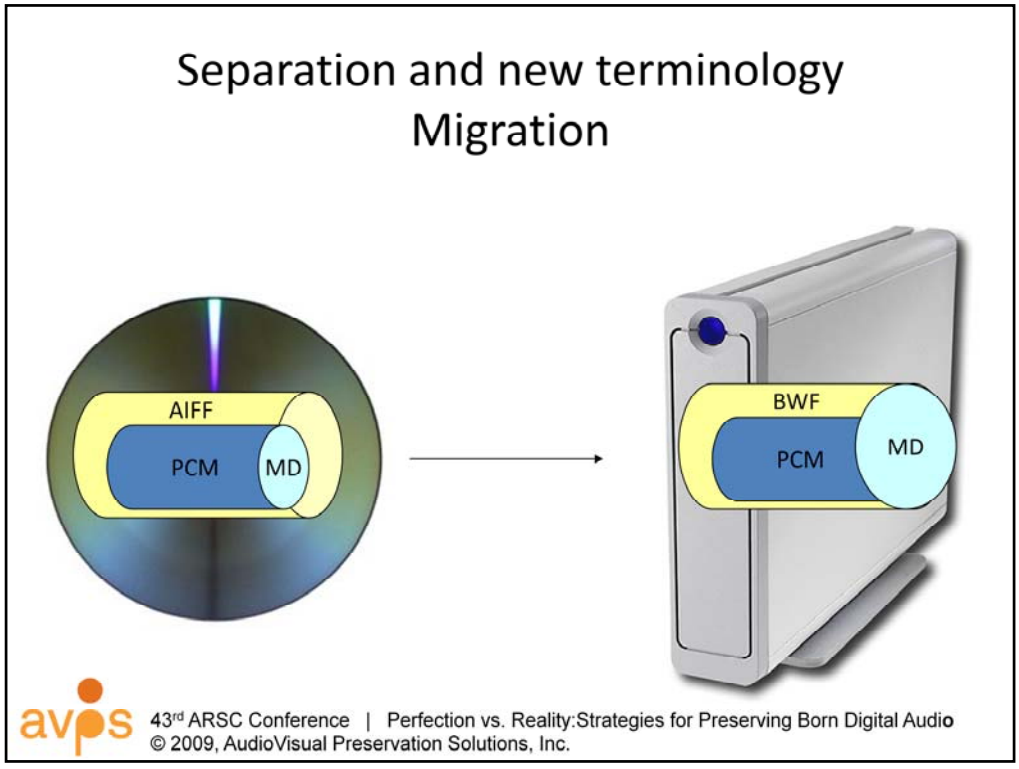
Refreshing is the act of copying data from one piece of media to another. This counters obsolescence and degradation of the media. Both of these are examples of refreshing data.

Separation and new terminology Refresh





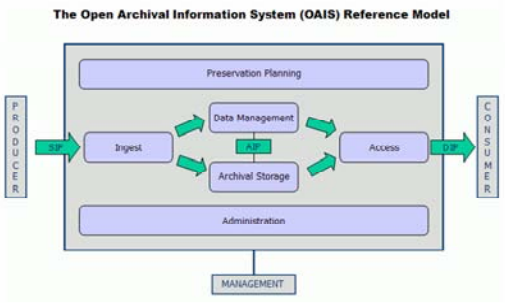
Migration is now the act of mapping or decoding and re-encoding the essence into a new codec and/or wrapper. This counters obsolescence of digital characteristics. Both of these are examples of data migration.



This is also migration. This example demonstrates a significant change in both terminology and interpretation of a principle.

Practices

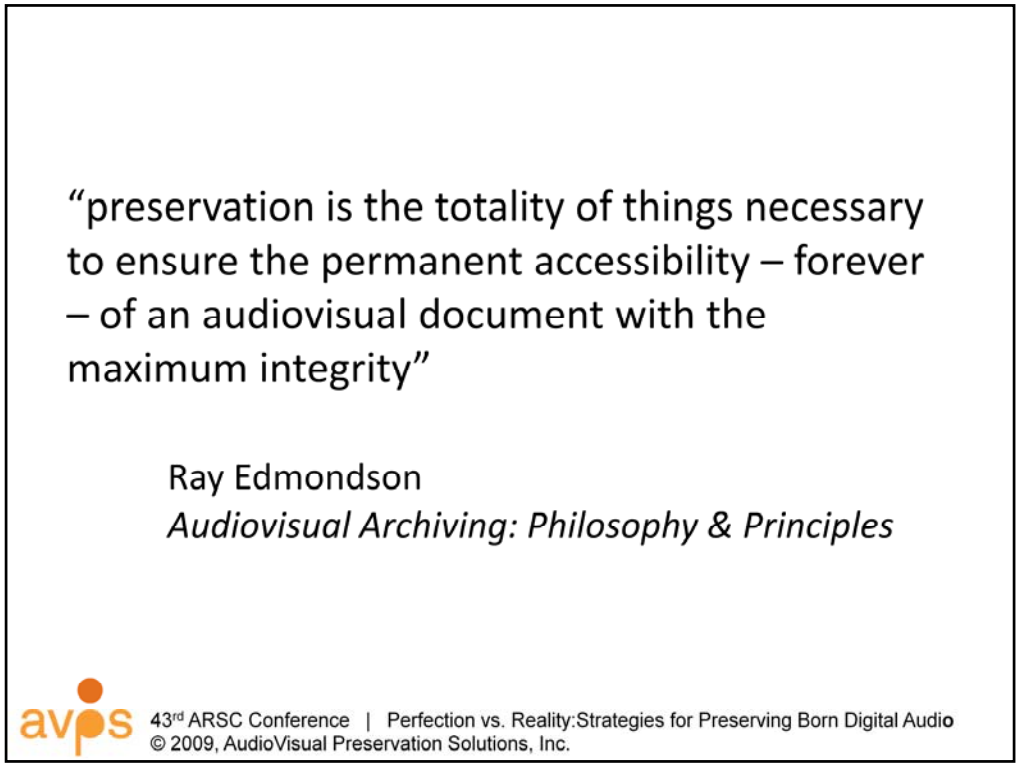
- OAIS Functions
  - Ingest
  - Store
  - Administrate
  - Maintain/Manage
  - Access/Disseminate



Now, shifting to practices and skill sets, I think it's ever important to reference OAIS, which often gets referred to in association with file-based content and digital repositories, but almost never gets referenced specifically in regard to physical archives.

This may give the impression that OAIS should only be considered for digital collections, but this is patently untrue. Rather, a lack of resources and expertise in archives disallowed us as a community from performing these activities with moving image and sound in the physical domain, but they were no less important 20 years ago than they are today. The opportunities that exist in the file-based domain in the way of convergence with information technology and expertise, automation, and increased access leading to increased awareness led to the uptake and enthusiasm of OAIS along with file-based archives because it became a widespread reality for the first time.

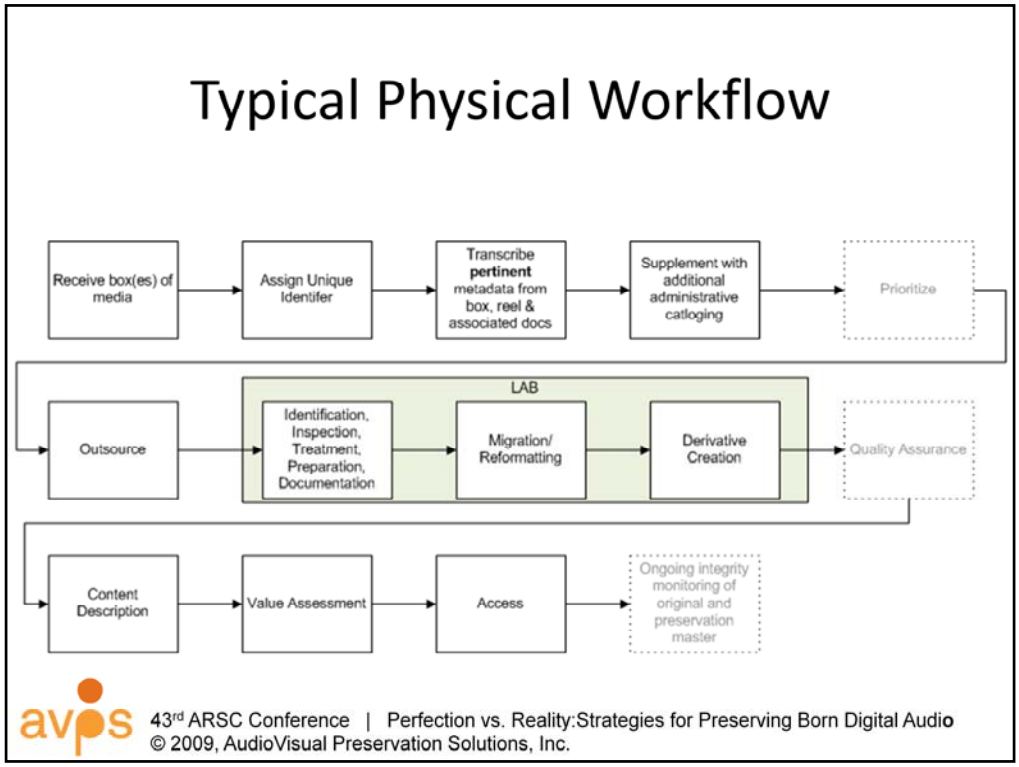
It is important to remember that just like our physical collections, our file-based collections require more than storage alone, even in the best of storage conditions. Preservation was never really and will never really be about the ultimate archival preservation format or the ultimate archival storage medium. Instead, it is about principles and strategy.



And as Ray Edmonson so aptly states "the totality of things necessary to ensure the permanent accessibility – forever – of an audiovisual document with the maximum integrity".

But it's one thing to look at the OAIS functions and to consider the "the totality of things necessary" and it's quite another to interpret and implement them in an ever-changing technological landscape with real world constraints such as resource limitations and limited time. Let's look at some real world aspects and some of the differences between the physical domain and the file-based domain.

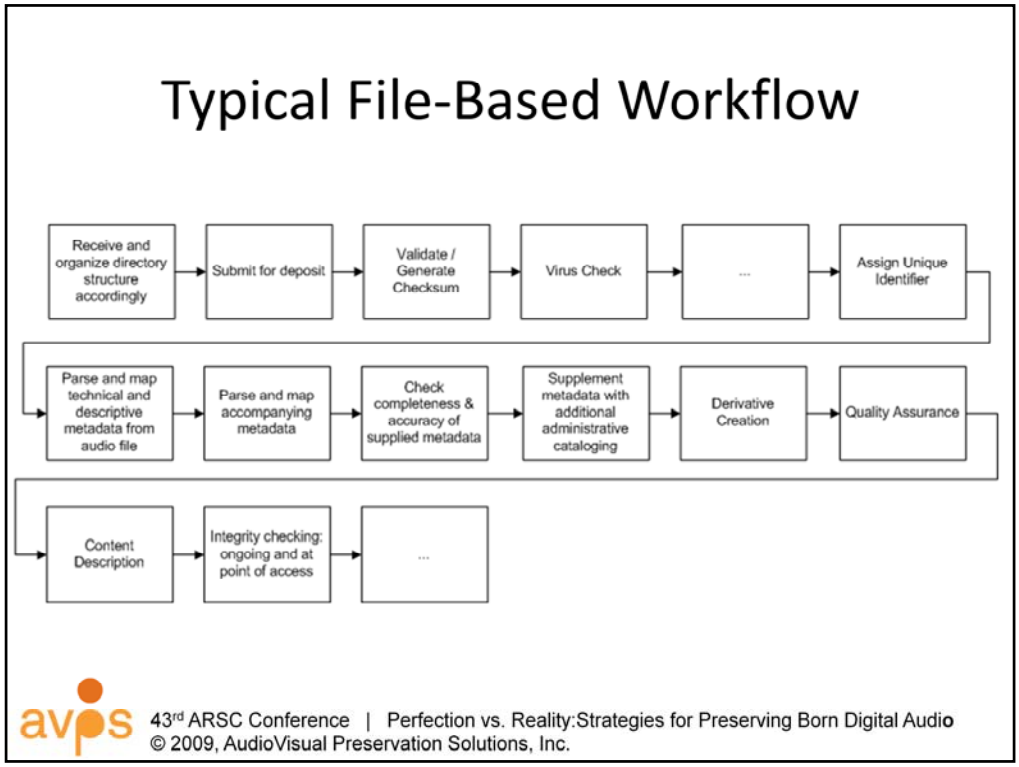
As a jumping off point to get into differences we'll start with one fundamental similarity. Accessioning physical and digital-born collections is similar in that archives generally have no control over the formats they receive. The quality, condition, resolution and so on is what it is. This causes major administrative problems – namely a wide variety of formats each with distinct technologies and associated risks that are difficult to manage for preservation and access.



In the physical domain this problem has been exaggerated due to what we see here in this typical physical workflow, where 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 would be 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, we all know that these activities usually take place over a very long time span in the physical domain.

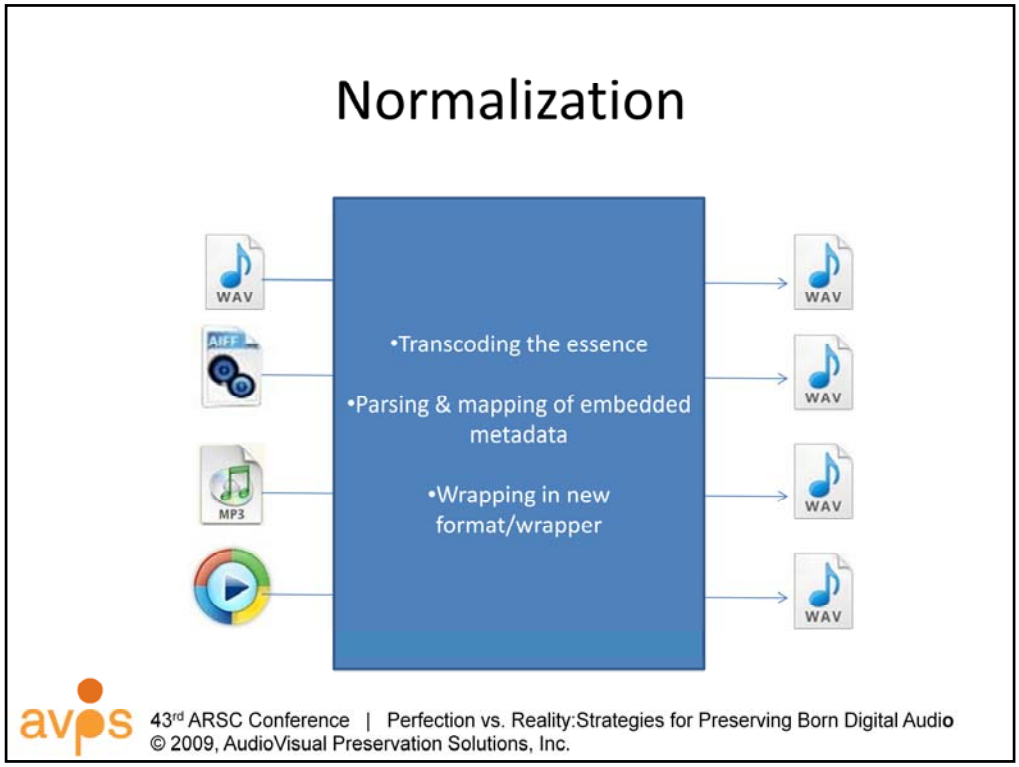
This lack of intellectual control prior to migration has been crippling to many organizations. It has led, and continues to lead to the inability to effectively prioritize, budget and allocate resources. It is not uncommon for funds to be allocated toward migration of content prior to there being real knowledge about the content, its value, the media condition or its uniqueness. Given the amount of content we have to preserve, think about the real impact of that for a moment. It's a stark reality. Fortunately there are some positive movements on this front in the way of education, innovation and more.



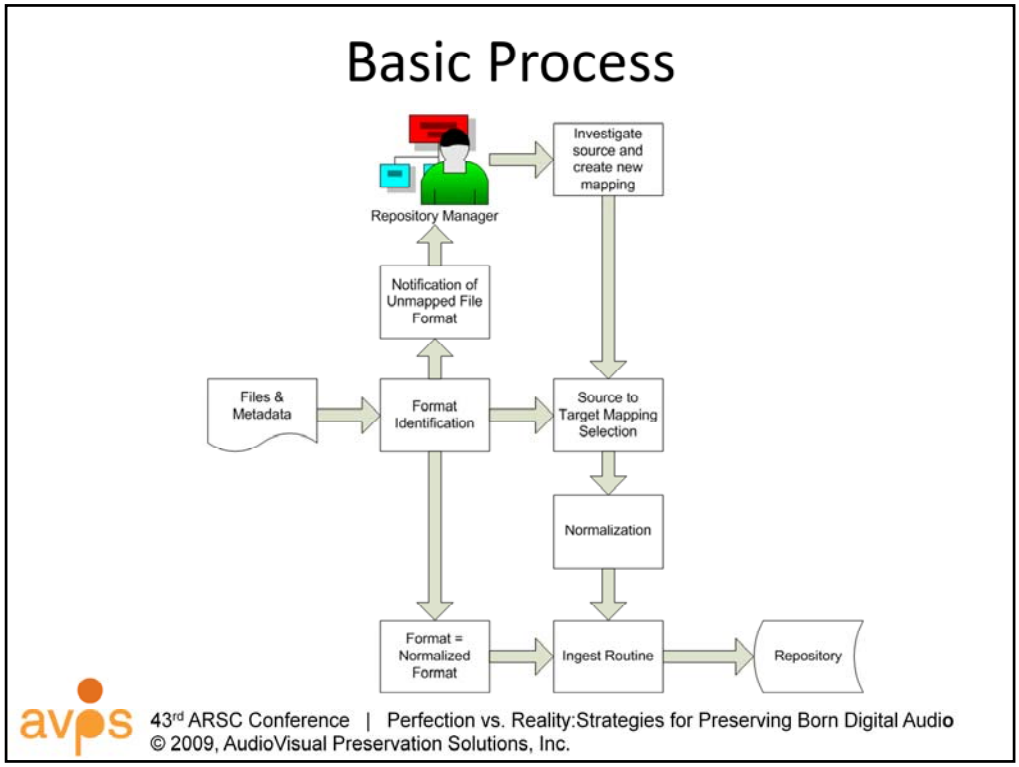
On a brighter note, with the move to the file-based domain, and the imperative of an actively managed environment a change was not only required - but possible due to the availability of a new toolset.

This image shows a typical basic file based workflow. In contrast to the long timeline necessary for the physical workflow, the timeline associated with this workflow could almost be considered negligible.

There are so many opportunities to compare these workflows and to look at how preservation principles are interpreted into practices differently, but the areas marked with the three dots are where I want to focus today. The activities represented by these dots are Normalization and Obsolescence monitoring. Not new concepts, but practically and technically implementable for the first time.

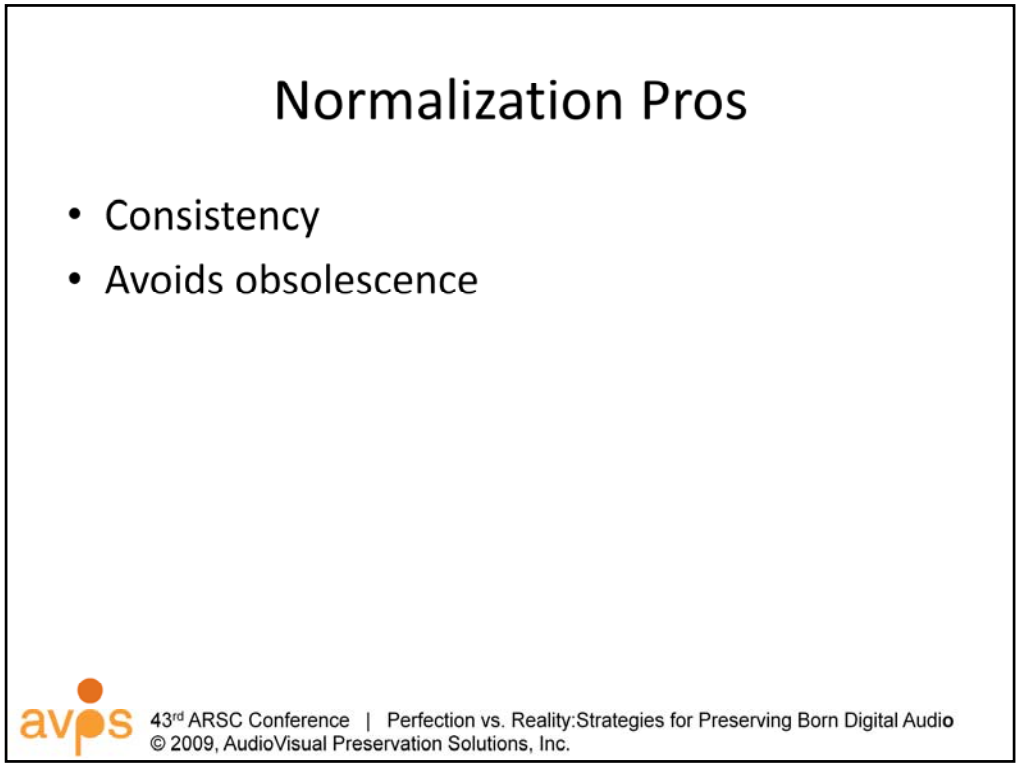


We will start with Normalization. Normalization takes place at the point of ingest and is essentially proactive migration. It is not dissimilar in theory (and I stress “in theory”) from reformatting a variety of analog assets to one common file format specification. At the point of ingest files are trans-coded to one consistent file format.



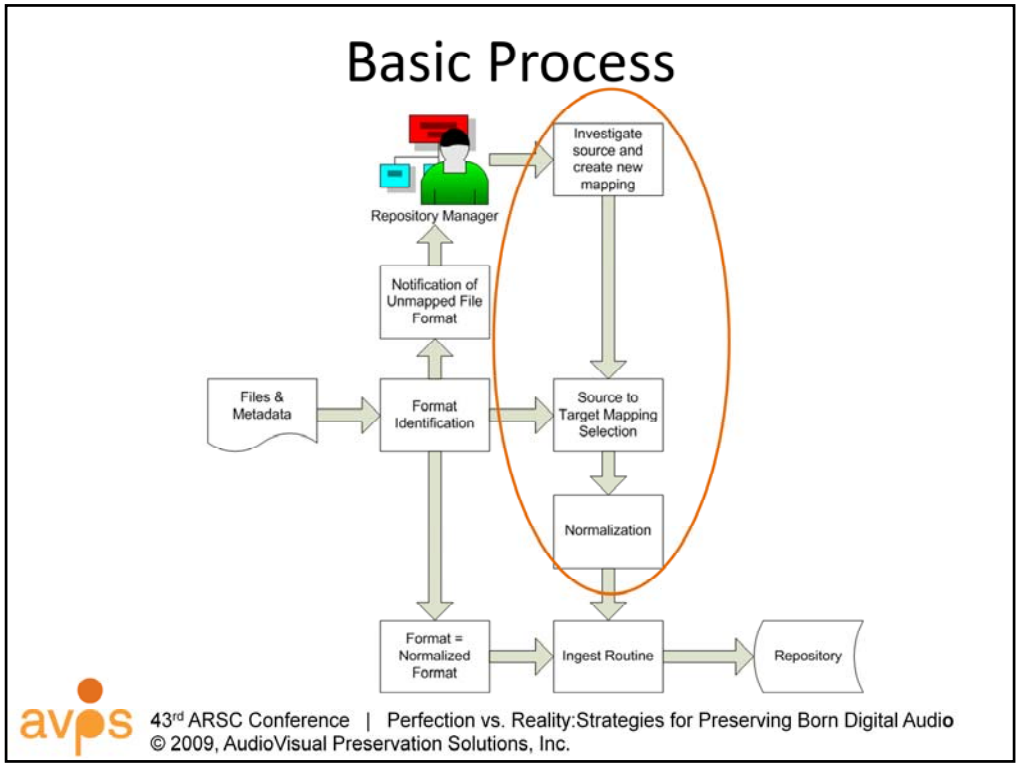
The basic logic of implementation looks something like this.

While this looks good on paper from a high level perspective, technically and practically it presents some significant issues, with only a couple real major advantages.

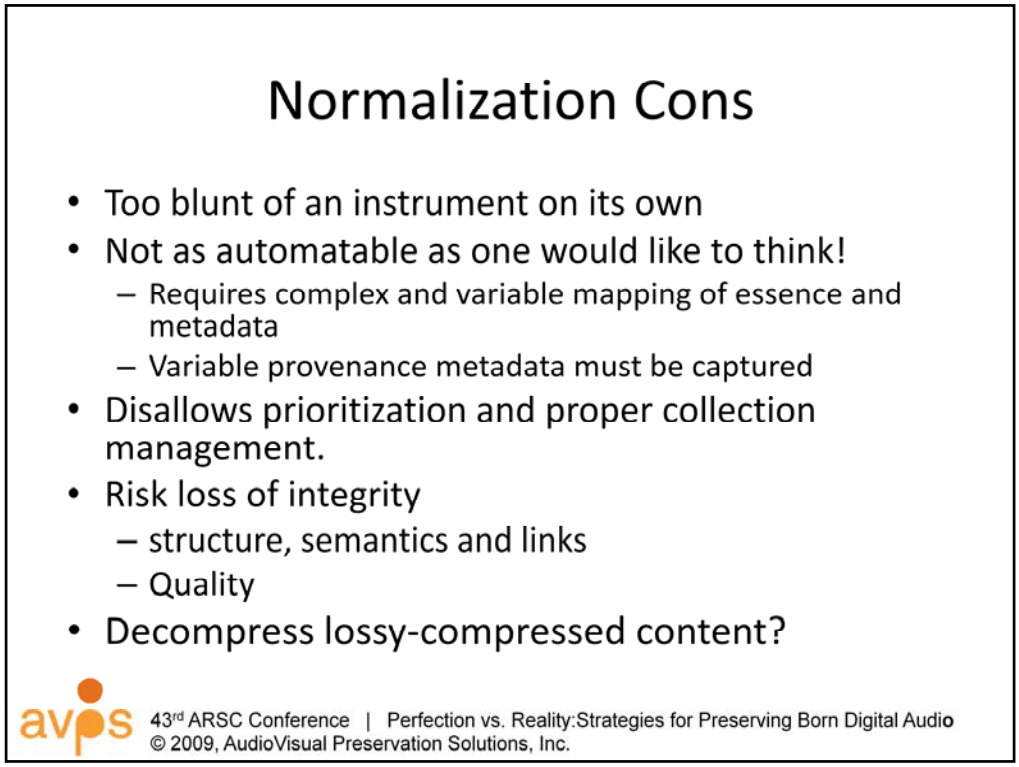


Consistency, is a great friend to preservation. It is good not only for immediate and near term collection management, but especially for the next migration which will take place. It is a nice ideal…

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



The major practical issues come into play in the activities highlighted here. The reality is that with variable input of codecs, wrappers and metadata the task of creating the trans-coding 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.

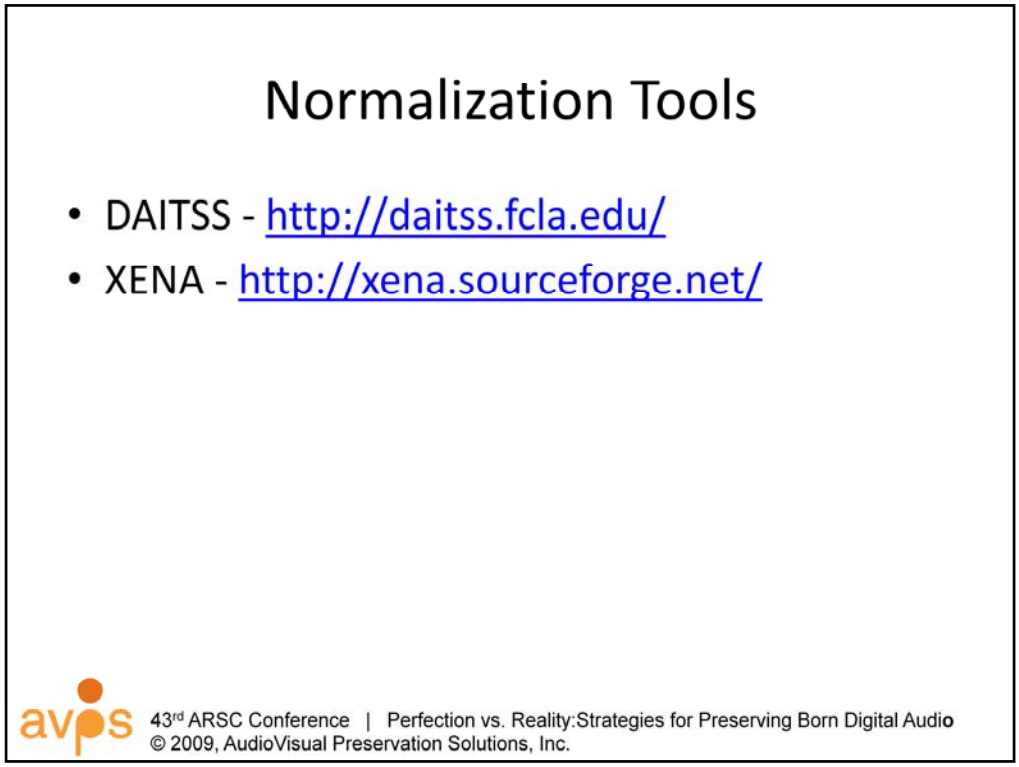


Ultimately, the cons are many. First, after scanning down the list of cons you will notice that as a sole tool, it's too blunt of an instrument. The tasks mentioned in the last slide as being challenging are time consuming and resource intensive unless there is a high level of consistency on input.

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 trans-coding and mapping, creating a low quality output as our master.

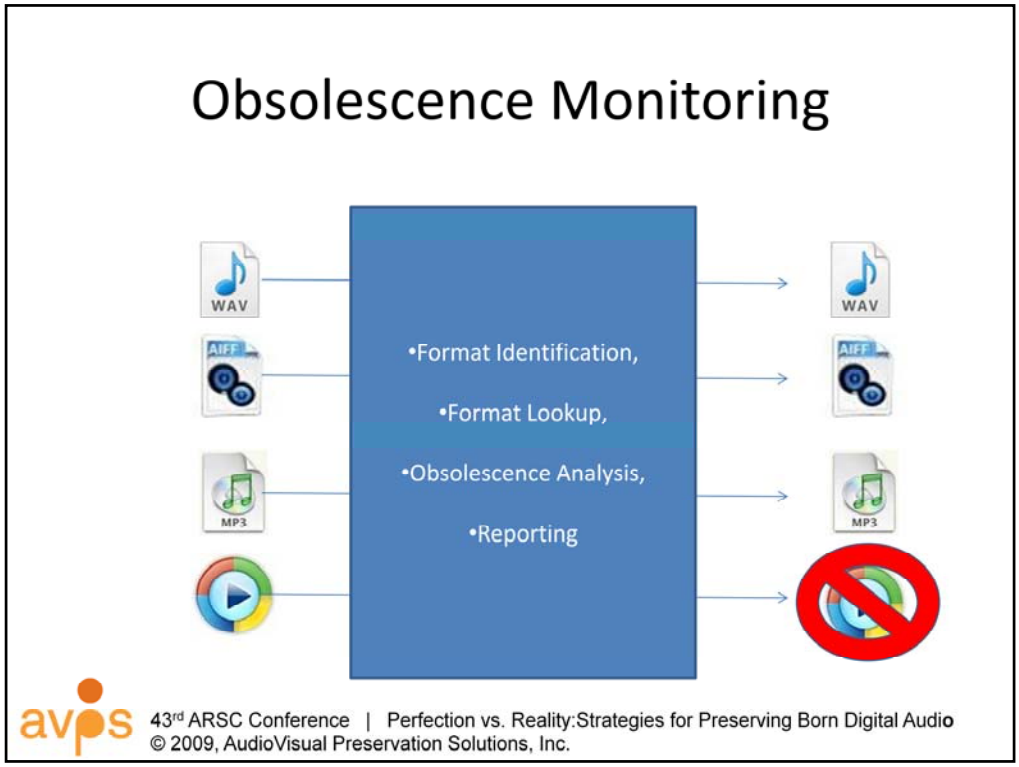
Lastly we see a practical consideration of storage. The likely candidate for a normalized output would be uncompressed, while many of the input formats would likely employ lossy compression. The result is the worst of both worlds. The negative of a larger file coupled with the negative of poorer sound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

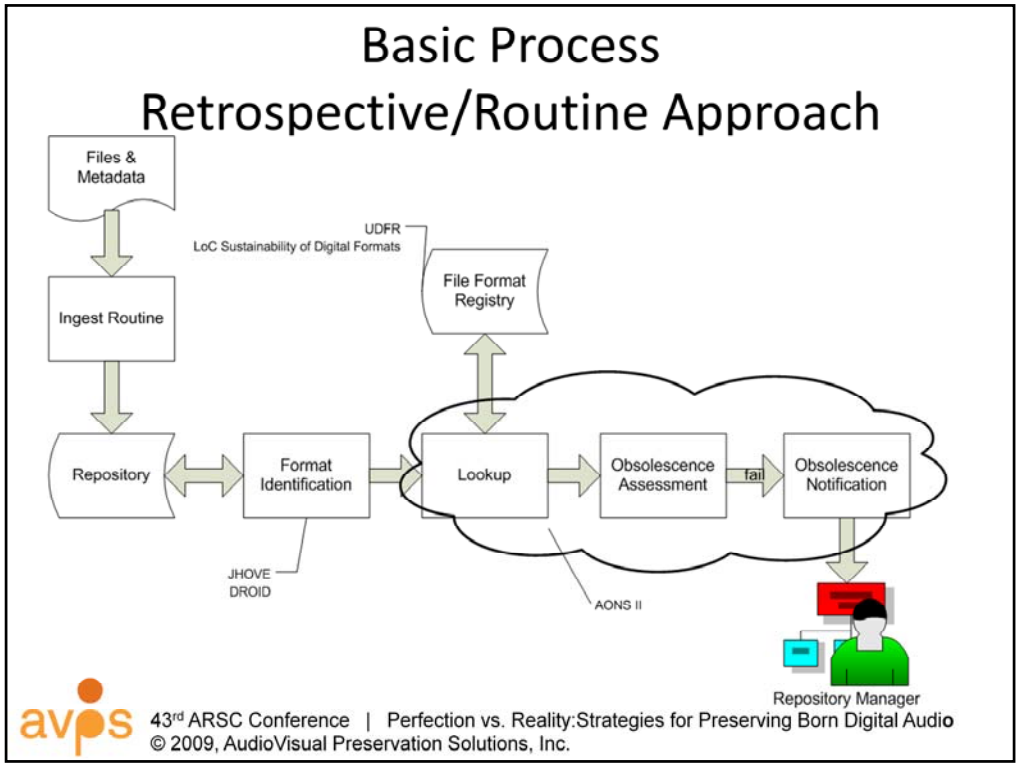
For smaller collections or organizations without access to the appropriate expertise, there are applications with GUIs (Graphical User Interfaces) that can be used in a batch mode to perform some of these activities, although you will hit an obstacle when looking for metadata mapping capability.



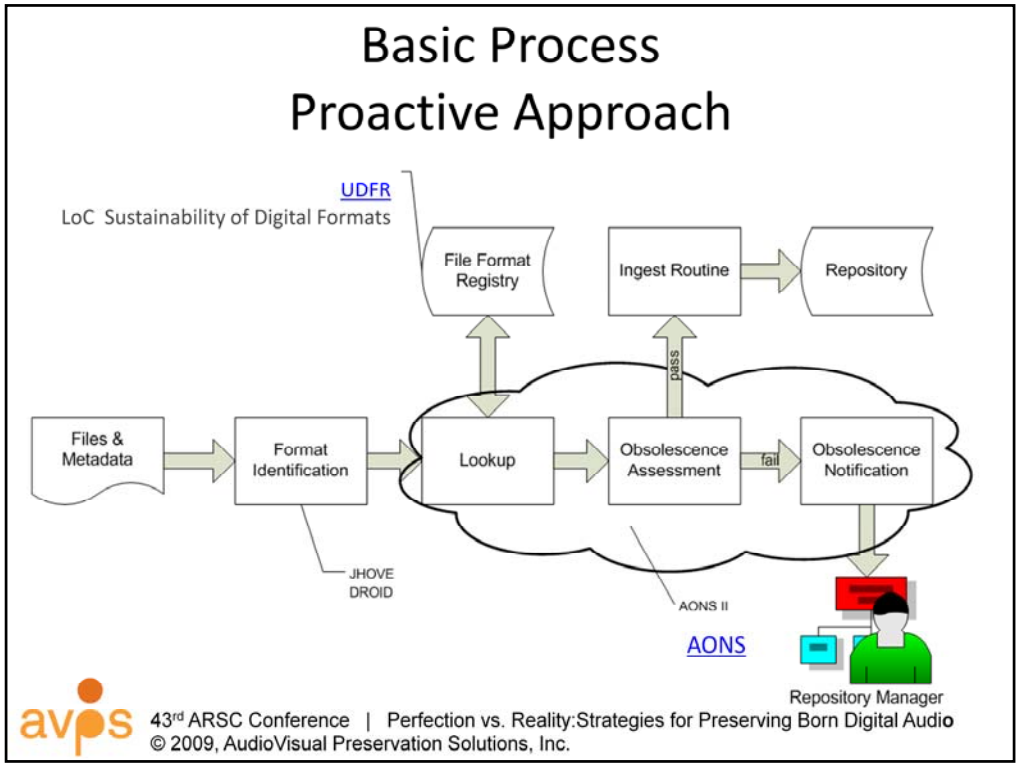
Obsolescence monitoring is a more sophisticated and complex mechanism, used at the point of ingest

- 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.



This is a look at the basic workflow associated with performing obsolescence monitoring retrospectively or on an ongoing routine basis.

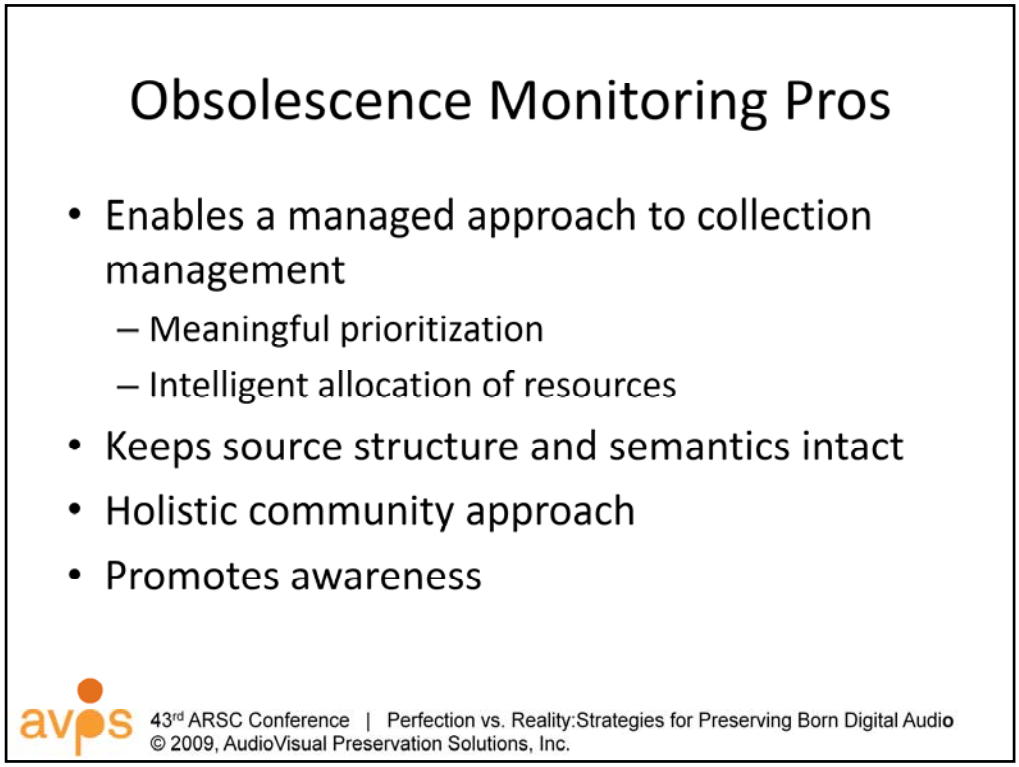


In this example we see obsolescence monitoring taking place at the point of ingest. In both of these examples you probably noticed some notes such as UDFR, LoC Sustainability of Digital Formats, JHOVE, DROID and AONS II.
These are examples of application types that are required to perform the tasks of Format Identification, Providing Format Lookup information, Obsolescence Analysis, and Reporting

First JHOVE, which stands for JSTOR/Harvard Object Validation Environment or DROID, which stands for Digital Record Object Identification, can be used to identify and validate the file format type. After the type is identified AONS II which at its core is an obsolescence assessment and reporting tool, Looks up the file type in a file format registry such as UDFR or the Library of Congress Sustainability of Digital Formats Registry. UDFR stands for Universal Digital Format Registry, and combines the efforts of two previous prominent initiatives known as PRONOM and GDFR.

UDFR and the LC site serve as format registries, documenting detailed information regarding things like external dependencies, rights issues, compression, standardization, documentation available, and so on.

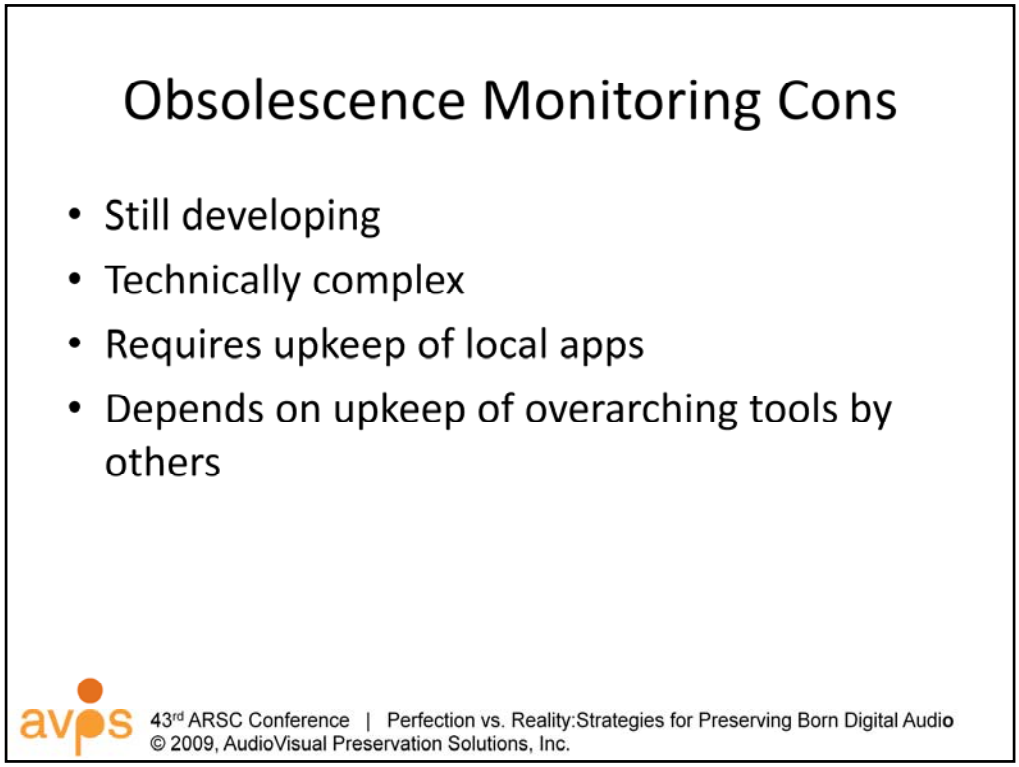
AONS II receives this information and analyzes the results according to criteria established by the AONS team. The result is an obsolescence rating. If it is poor it sends a notification to the Repository Manager. If it is low risk the file proceeds through the ingest process and is deposited in the repository.



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.



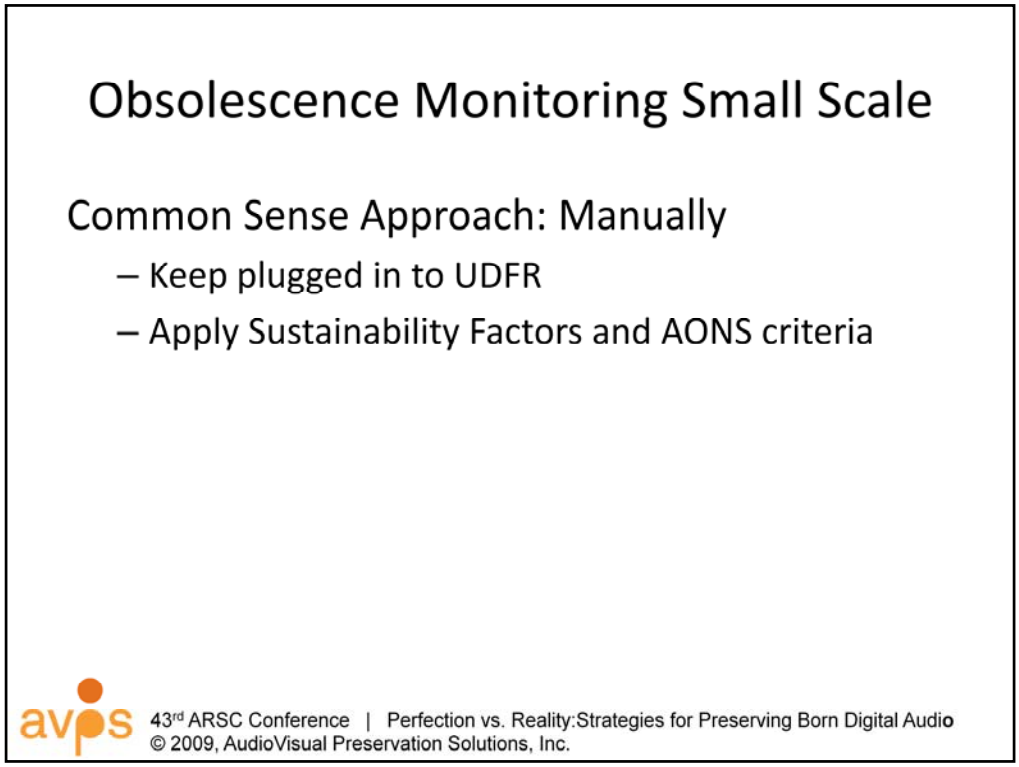
The main con is that these tools are still very much in development. The integration of the tools and their exchange of data is in need of refining in order to function more smoothly.

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

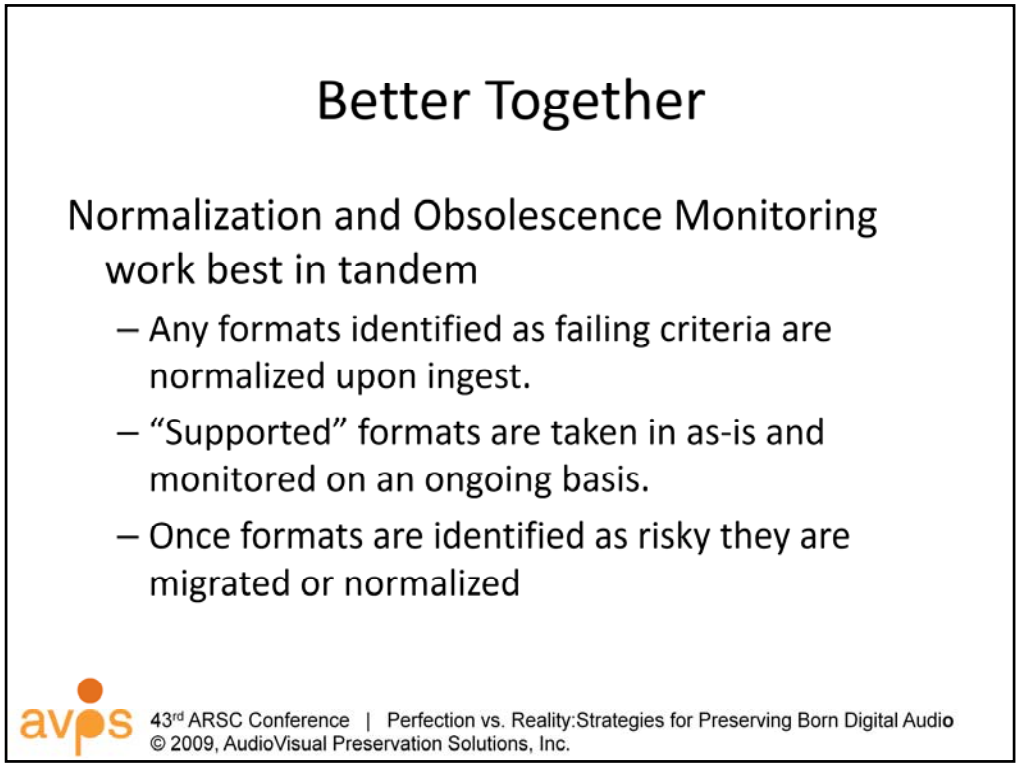
And as stated previously, this system heavily depends on others which can be both a good and a bad thing.

Obsolescence Monitoring Tools

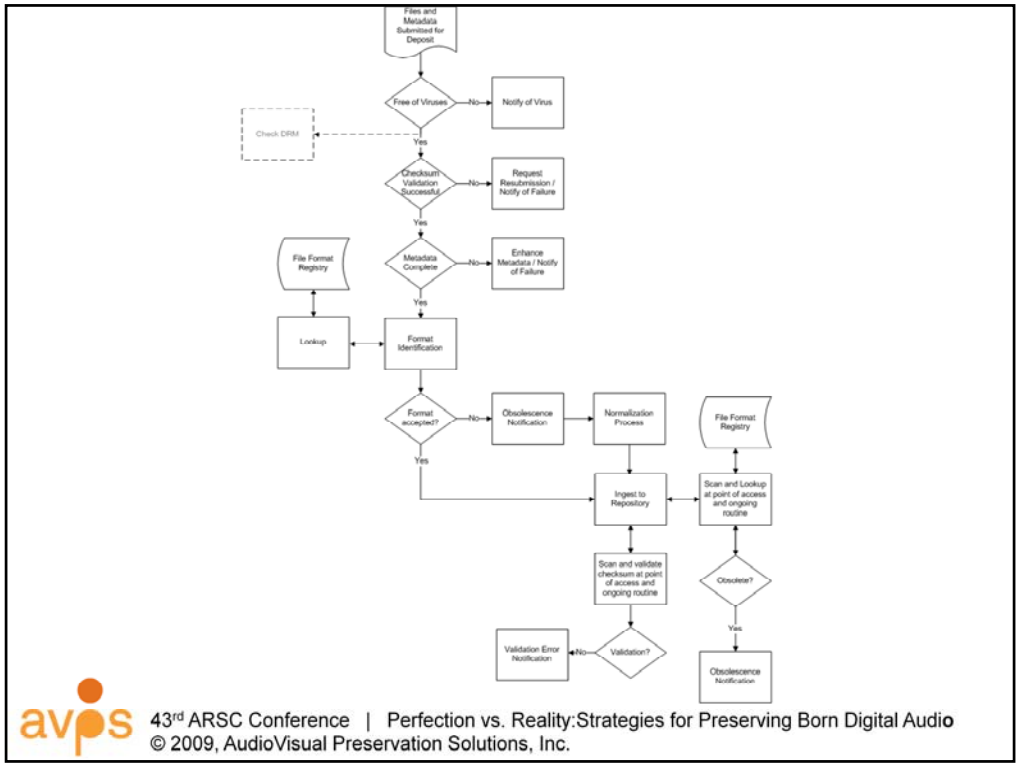
- AONS II (Automatic Obsolescence Notification System)
  - Download: http://sourceforge.net/projects/aons/
  - Blog: http://aons2dev.blogspot.com/
- UDFR (Unified Digital Formats Registry) - http://www.gdfr.info/udfr.html
  - PRONOM - http://www.nationalarchives.gov.uk/PRONOM/Default.aspx#
  - GDFR (Global 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



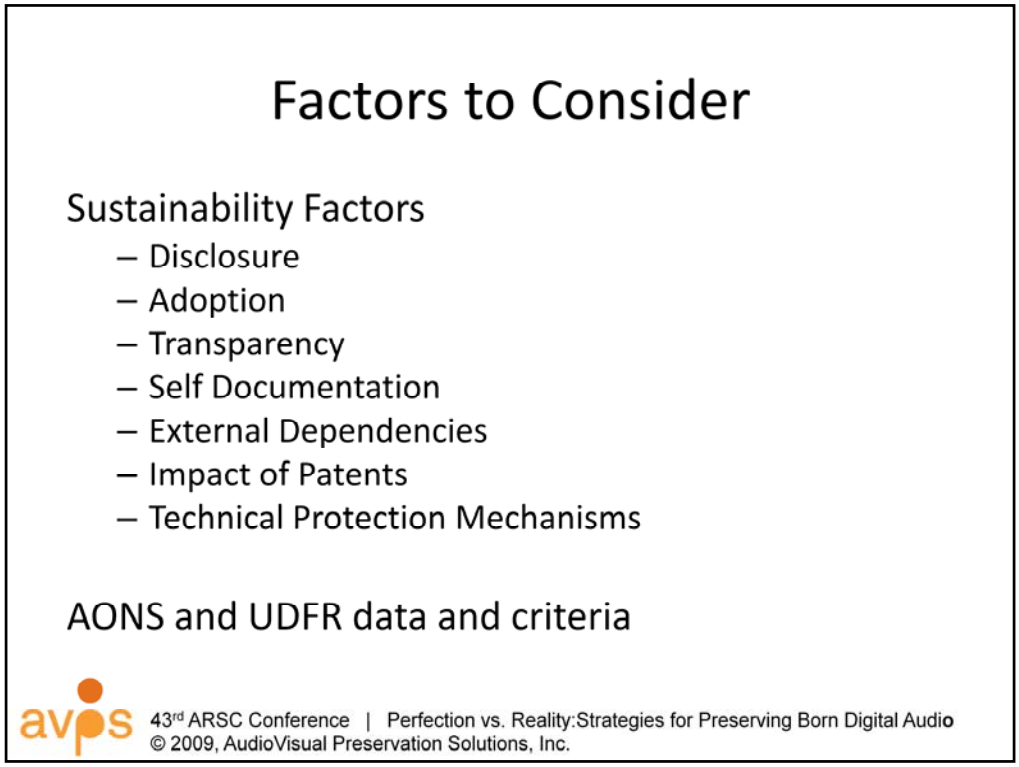
Recognizing the complicating factors just stated, obsolescence monitoring can also be performed on a small scale, or in situations where there is not too great a variety of inputs using some of the same tools manually. Less than ideal, but better than nothing.



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.



This decision tree and workflow looks at some salient ingest processes, utilizing both obsolescence monitoring and normalization.



Here are some factors and criteria to consider in deciding when to use obsolescence monitoring and when to use normalization and/or migration in your organization. These are the same criteria that the automated analysis reviews, but is relatively easy to understand and apply in cases where tools do not fit your organization or are not developed yet to meet your needs. As stated previously, these are good criteria to use when assessing any technologies including databases, asset management systems, backup software and hardware and more.

Other Factors to Consider

What type of organization and input?

- Production Centric = Potential control over deliverable to the archive
- Uniform input = pseudo control
- Variable input = no control

Other Factors to Consider

What your internal systems support

- formats
- bandwidth/resolution
- Archival systems vs. production and access systems

Some Examples of Actual Approaches

- Florida Digital Archive Obsolescence Planning- http://www.fcla.edu/digitalArchive/formatInfo.htm
- University of Minnesota Digital Conservancy Obsolescence Planning
  http://conservancy.umn.edu/pol-preservation.jsp#level2


avps
permanence persistence preservation
Chris Lacinak
AUDIOVISUAL
PRESERVATION
SOLUTIONS
917.548.8632 / f 866.264.4275
chris@avpreserve.com
www.avpreserve.com
350 7th Avenue Suite 1603
New York New York 10001
avps
43rd ARSC Conference | Perfection vs. Reality:Strategies for Preserving Born Digital Audio
© 2009, AudioVisual Preservation Solutions, Inc.